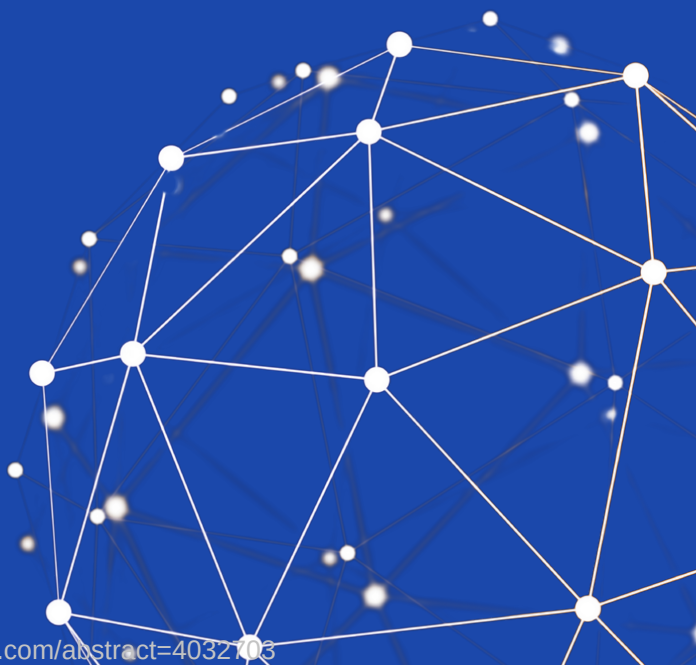
IS FACEBOOK KEEPING UP WITH INTERNATIONAL STANDARDS ON FREEDOM OF EXPRESSION? A TIME-SERIES ANALYSIS 2005-2020

KONSTANTINOS STYLIANOU
University of Leeds

NICOLO ZINGALES
FGV RIO

STEFANIA DI STEFANO
Graduate Institute, Geneva



Feb 2022

The authors wish to thank Nick Queffurus for research assistance and Paddy Leeversen for valuable comments. All errors remain the authors' responsibility. This project was funded through a competitive bid as part of Facebook's Research Grant programme. The grant was structured as unrestricted gift and no editorial or other control by Facebook was exercised over the project.

TABLE OF CONTENTS

PREFACES

Social media companies such as Facebook have been facing considerable criticism in recent years. They have been deemed responsible for spreading disinformation, inciting hate speech, discrimination, and violence. They also play a role in all sorts of electoral processes and referenda. Social media have evolved into a new form of 'public sphere', and the content moderation policies of social media platforms have a critical impact on the exercise of the rights to freedom of expression and information of their users. At a time when debates on freedom of expression and social media platforms are in the spotlight, this study asks whether criticism aimed at social media companies such as Facebook are legitimate and fair, and whether these companies can and should do more to align their content moderation policies with international human rights standards on freedom of expression.

Drawing from a vast range of international human rights law instruments and interpretations, this study aims to understand to what extent Facebook's content moderation policies are aligned with international standards on freedom of expression. By breaking down freedom of expression in different areas, the study shows how this freedom should not be conceived as a simple standalone right but is interrelated to many others.

The study reveals that Facebook was late in integrating international standards into their policies on content moderation. It then focuses, however, on the parallel evolution of both Facebook's policies and the evolving dynamics related to the international principles related to freedom of expression, the right to privacy and the right to information.

Distinctions between the 'public' and private' have for too long hampered clear thinking about how best to protect human rights. The private sphere has for too long been considered off-limits for regulation by human rights norms. Now that private companies such as Facebook clearly dominate the public sphere in so many ways, the time has come to set out a vision for how to ensure that private companies understand and recognize their human rights obligations and that there are effective remedies for those whose rights are violated. This study is not hampered by traditional assumptions about the public/private distinction, and after examining the situation in some detail, it goes on to make a series of practical recommendations for ensuring that social media companies, such as Facebook, fulfil their human rights responsibilities.

Andrew Clapham

Professor of International Law

Graduate Institute of International and Development Studies, Geneva

There are a number of important tasks in understanding how private actors, especially big platforms, develop and apply private rules in online spaces. The speed of change of these rules, the vast number of content moderation decisions that are taken daily, and the interaction between rules and algorithmic moderation practices make this a very challenging endeavour.

The role of platforms has revolutionized online communication spaces. Though the authors to the present study quite rightly describe the normative development online as a co-evolution of public and private rules, it is, arguably, a revolutionary co-evolution.

The authors elegantly show how platforms have amassed substantial power through their rules and how these develop in tandem with, and in contrast to, international standards. The authors rightly point to the intricacies of the relationship between private terms of service and public laws. Of particular interest to readers will be the comprehensive criticism of the interaction regime. The study finds disconnects between the international system for the protection of freedom of expression and correctly identifies power differentials between users and platforms.

What makes this study particularly valuable, is the nuanced treatment of the interaction of Facebook's standards and of global rules. Rather than criticizing the platform outright, the authors carefully analyze that depending on the subject area, Facebook's standards can be higher or lower than international standards demand. The authors offer robust criticism of the platform's approach to reinstatement and redress policies, categorical bans, chilling effects of overblocking, and protection of minors. Important developments, like the use of (and limit to) AI-based content governance are also discussed.

While focused on one platform, this study is an important analysis that provides a deep understanding of the challenging interaction of global and national public and private rules, in the normative order of one key private communication actor. It is essential reading, especially as Europe's normative approach to platforms matures. That some of the study's recommendations have already made it into the Digital Services Act, shows just how topical and convincing the authors have managed to make their analysis.

Matthias C. Kettemann

Professor of Innovation, Theory and Philosophy of Law

Department of Theory and Philosophy of Law, University of Innsbruck (Austria) / Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg

Considering the relevance of digital platforms for contemporary social interaction, especially the importance of Facebook with around 2.8 billion monthly active users, this report brings an invaluable set of findings by analyzing legal and contractual standards of Facebook terms of use vis-à-vis international human rights law. It is not an easy task to combine directions that were born and developed in parallel, such as, for instance, the principles of North American intermediary liability law that provided the foundation for Facebook´s terms of use, and States responsibilities under international human rights.

The researchers responded to this challenge by posing the right question: to what extent have Facebook´s content policy aligned with international standards on freedom of expression over the last fifteen years of its existence? The relevance of this question is evident in a context where authoritarian populist leaders are disputing the very notion of free speech in favor of discriminatory, misogynist and often violent discourses that are disruptive of democratic rule of law.

By comparing the development of standards and policies adopted by Facebook, including the establishment of its Oversight Board, with the guidance of the international community, the researchers found that their movement is slow and less efficient compared to society's urgent needs, the fast development of new technologies and the international community's agenda. New social movements, contemporary human rights blueprints, and enforcement of corporate responsibility based on the UN Business and Human Rights principles expect Facebook and other social media companies to step up and play a proactive role to in defense of users' freedoms of expression. One could even argue that Facebook has a greater role to play, due to the large scope of its influence: defending democracy itself.

At the same time, the study shows that the international community, UN bodies and national states are timid in proposing policies and new legislations that engage with non-state actors such as Facebook, which are playing a pivotal role in shaping the public sphere debate in several areas. This is not new in human history: life flows more rapidly than the process of law formation, which requires the adoption of international principles before they can feed into the content moderation policies of companies like Facebook and guide responses to new situations that could harm freedom of expression. Meanwhile, social media companies like Facebook are confronted with many interpretative challenges: the possible conflict between the law of the land, the law of the platform and international law; cyberspace sovereignty, diversity within and across jurisdictions on substantive norms, remedies, procedures, state and corporate responsibilities, accountability of artificial intelligence use, and so on.

First and foremost, the main task of our generation is to defend freedom of expression as a means to advance equality and non-discrimination. The report brings to our attention the unbalanced game of power amongst superwealthy platforms, communities and individuals, and the criticality of various forms of expression and participation, especially when we consider the use of AI and its negative impact over racism and xenophobia. It also shows efforts from the US and EU legislatures to act on such matters, but there

are many countries struggling to protect freedom of expression of its nationals – particularly in the Global South – with weaker effects over such corporations due to legal and political constrains.

I want to conclude by congratulating the authors on this excellent piece of research and by raising an uncomfortable question: considering Facebook´s moderation decisions in the context of the attacks at the Capitol Hill on Jan 6th, 2021, how would its response have been if the same event occurred in the presidential election of a country in the Global South, such as Brazil? Again, history unfolds at a faster pace than law-making; by adhering to human rights law and principles, Facebook can effectively support freedom, equality and democracy.

Denise Dourado Dora

Executive Director of Article 19 Brazil

PART 1: CONTEXT

1. Introduction and Summary

Facebook has been a common target of criticism for its policies and practices on various fronts. In 2018, Wired magazine even published a year-in-review list of the "21 (And Counting) Biggest Facebook Scandals",[1] and in 2021, the Wall Street Journal compiled "The Facebook Files" documenting how Facebook "knows, in acute detail, that its platforms are riddled with flaws that cause harm, often in ways only the company fully understands."[2]

Many of those points of criticism revolve around Facebook's core moderation conundrum, namely what content to allow on its platform and what content to ban. This question emerges in various forms and in different contexts, be it the regulation of a type of speech (e.g. hate speech), the protection of a class of Facebook users (e.g. children), the chilling effects policies may have on freedom of expression on Facebook in general (e.g. Facebook's relationship with governments), or the effect certain of its policies, such as fake news and misinformation, can have well beyond its platform to include effects on democratic discourse, public health, and other areas of general interest. With a monthly active userbase nearing 3 billion,[3] Facebook serves as the world's "public sphere" — a virtual place where people connect and develop a common polity, or at least a common base,[4] and it is therefore not hard to see why its policies exert such great influence across the world.

---

[1] Issie Lapowsky, 'The 21 (and Counting) Biggest Facebook Scandals of 2018' (*Wired*, 12 December 2018) https://www.wired.com/story/facebook-scandals-2018/.

[2] 'The Facebook Files' *The Wall Street Journal* (September 2021).

[3] 'Number of Monthly Active Facebook Users Worldwide as of 2nd Quarter 2021' (*Statista*, July 2021) https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.

[4] Dominik Batorski & Ilona Grzywińska, 'Three Dimensions of the Public Sphere on Facebook' (2018) 21 Information, Communication & Society 356.

Despite (and perhaps also thanks to) the relentless criticism, Facebook's content policies have evolved dramatically since their first version in 2004. The introduction of the Community Standards in 2010 in particular, where Facebook details how it handles different types of content, has been instrumental in clarifying the scope and limits of users' freedom of expression on the platform, and the creation of the Oversight Board in 2020 has been a global first in accountability and in instituting a quasi-judicial review of Facebook's policies.[5]

Are Facebook's efforts in vain then? Is the criticism around Facebook's content policies fair and justified? One way to approach the criticism is to blame sub-optimal policies on the complex nature of content moderation.[6] Because free speech "is the matrix, the indispensable condition of nearly every other form of freedom"[7] its boundaries are shaped by a dizzying array of considerations, which is only compounded by the fact that Facebook operates on a global scale. Given the immense complexity of freedom of expression regulation, it is next to impossible to strike the right balance, and complaints will always persist. In fact, Facebook suggests that "one of the biggest issues social networks face is that, when left unchecked, people will engage disproportionately with more sensationalist and provocative content. This is not a new phenomenon. It is widespread on cable news today and has been a staple of tabloids for more than a century. […] Our research suggests that no matter where we draw the lines for what is allowed, as a piece of content gets close to that line, people will engage with it more on average — even when they tell us afterwards they don't like the content."[8] This inclination of people to engage more with borderline acceptable content (however it is defined) makes the design of content moderation policies even more challenging.

---

[5] Kate Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2019) 129 Yale Law Journal 2418. For criticism of the Oversight Board, see Dipayan Ghosh, 'Facebook's Oversight Board Is Not Enough' (*Harvard Business Review*, 16 October 2019) https://hbr.org/2019/10/facebooks-oversight-board-is-not-enough.

[6] Michael P. Zuckert, 'The Insoluble Problem of Free Speech' (*National Affairs*, Fall 2018) https://nationalaffairs.com/publications/detail/the-insoluble-problem-of-free-speech.

[7] *Palko v. Connecticut*, 302 U.S. 319, 327 (1937).

[8] Mark Zuckerberg, 'A Blueprint for Content Governance and Enforcement' (*Facebook*, 2018) https://www.facebook.com/notes/751449002072082/.

True as it may be, this approach is not particularly helpful or illuminating. While perfect content moderation policies may not exist, some are better than others. The more meaningful question is whether social media platforms like Facebook design their content moderation policies pursuant to the law, best practices, and guidelines available to them. If they do, they are at least doing the best they possibly can.

Considering Facebook's global scale, guidance from the international community would be the obvious place to start. A large body of international instruments, including binding treaties, covenants, and primary and secondary EU law, and non-binding guidelines, declarations, reports, and recommendations produced over the past 70 years, has developed the content and contours of freedom of expression, and, as they reflect global standards set by international organizations, they can be relied upon to establish policies that enjoy a high degree of legitimacy.[9]

Technically, corporations, like Facebook, are not directly bound by international law. However, as the UN Guiding Principles on Business and Human Rights (UNGPs) have clarified, they still have a responsibility to respect human rights throughout their business activities,[10] including the right to freedom of expression. This *responsibility to respect* stems from a near-universal recognition of social expectations in that direction.[11] This framing allows us to distinguish between the state duty to *protect*, which stems from the international obligations states have undertaken under international human rights law, and the corporate responsibility to *respect*, which is understood as a global standard of expected conduct.[12] In light of this, we see the rights, obligations, and accompanying interpretations enshrined in international instruments as best practices, guidance, and standards that private corporations can and should aspire to.

Moreover, the commentary to Guiding Principle 12 recognizes that although business enterprises can have an impact on virtually every human right, some human rights might be at

---

[9] Dataset available at https://doi.org/10.5518/1072.

[10] United Nations, Guiding Principles on Business and Human Rights (2011).

[11] John Ruggie, 'The Social Construction of the UN Guiding Principles on Business & Human Rights' (2017) HKS Faculty Research Working Paper Series RWP17-030 www.hks.harvard.edu/publications/social-construction-un-guiding-principles-business-human-rights, 13-14.

[12] Ibid 15.

particular risk in particular industries and contexts and, as such, they deserve heightened attention.[13] If the responsibility to respect human rights refers, at a minimum, to the rights enumerated in the International Bill of Human Rights and the principles concerning fundamental rights set out in the International Labour Organisation's Declaration on Fundamental Principles and Rights at Work,[14] the commentary also adds that in some circumstances it may be necessary to consider additional standards.[15] Given that freedom of expression is one of the rights at heightened risk in the social media industry, we can and we do hold corporations accountable to standards that go beyond the International Bill of Human Rights.

Recognizing the relevance of international human right laws, Facebook has opened itself up to them: its very own Oversight Board, in adjudicating content moderation disputes that arise between Facebook and its users, is required to "pay [...] attention to the impact of removing content in light of human rights norms protecting free expression."[16] The Oversight Board is tasked with interpreting and enforcing Facebook's policies, and in the process of determining what the policies' limits and meaning should be, it looks at international human rights standards.

In this context, we set out to assess the compatibility of Facebook's content policies with applicable international standards on freedom of expression. We do so, not only regarding Facebook's current policies (as of late 2020), but historically as well starting from Facebook's founding. The historical dimension allows us to observe not only how Facebook's response has changed through time, but also how freedom of expression has evolved and how emphasis has shifted to new areas of speech, issues, or groups, particularly online. While the hard core of the right to freedom of expression has remained intact, the modality of online expression brings new challenges to the foreground. The parallel tracking of the evolution of Facebook's content policies and of freedom of expression standards, allows us to capture this cat-and-mouse game, and to assess the adequacy of both the international community's response and of Facebook's response

[13] Guiding Principles on Business and Human Rights (n 10), Guiding Principle 12.

[14] Ibid Guiding Principle 12.

[15] Ibid Guiding Principle 12.

[16] Facebook Oversight Board Charter 2019, Article 2§2.

to those changes. Our research aims to highlight areas where Facebook was quick or slow to adopt policies that reflect international standards, and to assess the adequacy of its "compliance."

We do not assess whether and how Facebook actually enforces its policies. Our focus is rather on the content of Facebook's policies, as they are reflected in the text of Facebook's Terms of Service, Community Standards and associated documents. These documents collectively form the legal contract between Facebook and its users, they define what users can and cannot do, and they spell out Facebook's commitments to users. Because the Terms of Service, Community Standards and associated documents are the only legally binding agreements between Facebook and its users, they are accordingly the only contractual basis on which users can rely to challenge Facebook's conduct toward them. It is important, therefore, to examine these documents, not only for the rights and obligations they include, but also for those they do not include. Any policy that is not included in the binding contracts is more of a best-effort, good-will gesture that users do not have an enforceable expectation that Facebook uphold.

A number of insights derive from our study.

- Our overall finding is that **in virtually all areas of freedom of expression we tracked, Facebook responded slowly to develop content moderation policies that were up to international standards. While the international community was more proactive, it too missed opportunities for timely guidance on key areas**. The freedom of expression areas we tracked included terrorism speech, hate speech, false news, nudity, bullying, minorities and protected characteristics, intellectual property limitations to freedom of expression, the role of anonymity, protections afforded to children, access to remedies, and the chilling effects Facebook's compliance with government requests for user data may have. We do note that over time Facebook's policies have become more elaborate, and more protective of user expression, but, given its "move fast and break things" capabilities, progress was often achieved with a considerable time lag from the point that guidance by the international community became available.
- We also find that the take-it-or-leave-it approach Facebook imposes on users regarding its content policies (not unlike other social media companies) disregards the well-established "legality, necessity, proportionality" standard in international human rights law, which allows for less restrictive rules on expression. We acknowledge that a

proportionality approach makes content moderation harder and less uniform than categorical bans, but our focus here is speech maximization, not administrability.

- On the other hand, we find that the international community has not engaged sufficiently with non-state actors, such as Facebook. We acknowledge that at the international level the laws, standards, and guidance are normally developed by states for states. However, as recognized in the UNGPs, corporations have a responsibility to respect human rights. This recognition has been a long-standing desideratum finally finding written standing in 2011, and yet our research indicates that the majority of international instruments are either not addressed at or do not account for non-state actors, despite recognizing their immense intermediation power and their role as private regulators of speech. Therefore, the international community has passed up opportunities to steer social media companies such as Facebook in the direction of the standards and policies they should adopt.
- We lastly observe that in 2009 Facebook instituted a process to involve users in the shaping of its content policies by giving them the right to vote on proposed changes.[17] This direct democracy measure was a significant experiment in enabling people to have a say in the rules that would go on to delimit their freedom of expression on Facebook's platform. While the experiment quickly failed, it also marked the first time a major global platform involved its users in its governance.

We conclude with a few recommendations on how Facebook can improve its content policies to safeguarding of users' freedom of expression on Facebook:

- We recommend that Facebook recognize and protect a content moderation acquis, meaning that any future changes to its policies will not weaken users' right of freedom of expression on Facebook. Going forward, any new limitations, clarifications, and modifications to the content policies must follow the "legality, necessity, proportionality" standard.

[17] Facebook, 'Facebook Opens Governance of Service and Policy Process to Users' (*Facebook*, 26 February 2009) https://about.fb.com/news/2009/02/facebook-opens-governance-of-service-and-policy-process-to-users/.

- We also recommend the general applicability of the proportionality standard, particularly where detailed guidance is missing. Any limitations must be proportionate, and categorical bans on certain types of expression should be extremely limited.
- Moreover, we recommend that Facebook include a description of users' access to remedies in its Terms of Service (ToS)/ Community Standards (CS), the only binding documents between Facebook and its users, to improve the legitimacy, accessibility and predictability of the mechanism. Facebook should further give users a right to explain why they think their content should be allowed on the platform.
- We also recommend that Facebook commit in its ToS/CS to providing an illustration of the AI mechanisms used in content moderation. Users should also be explicitly conferred a right to be informed about AI-driven adverse decisions such as removal of content and account termination, a right to receive an explanation for such decisions, and the ability to contest it with the involvement of a human reviewer.
- Lastly, we recommend a bona-fide exception for the use of non-authentic names, allowing users to provide reasons as to why pseudonymity might be necessary for them to exercise their freedom of expression right on-site.

The following Sections build our research case: In part I, Sections 2 and 3 discuss the rising power of platforms and their private ordering regime, i.e. the quasi-regulatory status they enjoy vis-à-vis their users' rights. Section 4 documents the various laws that impose moderation obligations onto platforms like Facebook or establish safe harbours that protect platforms from incurring liability for user-generated content. Section 5 connects content moderation policies with their enforcement and discusses the difficulties of making compliant but also administrable policies. Section 6 presents our methodology for collecting and analyzing the data underpinning our study, namely the body of international hard and soft law from 1948 to 2020, and the body of Facebook's content policies from 2004 to 2020. In Part II, Sections 1 through 5 provide contextual observations on how Facebook's content policies evolved alongside guidance by the international community. Section 6 contains the bulk of our analysis presenting in detail the key areas of freedom of expression, how Facebook responded historically to the emerging challenges, and the extent to which it incorporated available guidance by the international community. Section 7 presents our recommendations.

2. Platform power and the growing importance of the law of the platform

A wide range of interpersonal relations takes place every day on Facebook, which, with roughly 2.8 billion monthly active users[18]), is currently the largest social media platform worldwide. The sheer number of users, each with their own posts and interactions, provides a tangible demonstration of the prominent role that Facebook's platform has assumed in the fabric of society and social interaction. To give some order of magnitude, readership numbers at major news organizations pale in comparison, including for instance the Wall Street Journal with 2.35 million subscribers and the New York Times with 6.9 million.[19] Viewership numbers of US television channels are also nowhere near Facebook's user base, ranging for instance from 1.15 million at MNSBC to 4.5 million at Fox News.[20] Of course, such comparisons with traditional media make little sense from a competitive standpoint, considering that Facebook's experience offers much more than a channel for news distribution. But they do give us a sense of the unparalleled reach, and as a consequence, the enormous impact that Facebook has on public discourse with its rules on what content or activities are permitted on the platform.

Despite this unprecedented scale, Facebook and other social media platforms have been subject, at least until recently, to a lesser degree of regulatory oversight than traditional media organizations. This is precisely because of their nature as "platforms", or in their own words, "technology companies."[21] Unlike media, platforms do not actually produce content—that is

---

[18] 'Facebook: Number of Daily Active Users Worldwide 2011-2021' (*Statista*, 21 May 2021) https://www.statista.com/statistics/346167/facebook-global-dau/.

[19] Jeffrey A. Trachtenberg and David Marcelis, 'Publishers of Wall Street Journal, New York Times Ride Subscription Growth to Higher Profits' *The Wall Street Journal* (5 November 2020).

[20] Amy Waston, 'Leading Cable News Networks in the United States in May 2021, by Number of Primetime Viewers' (*Statista*, 07 June 2021) https://www.statista.com/statistics/373814/cable-news-network-viewership-usa/.

[21] Michelle Castillo, 'Zuckerberg Tells Congress Facebook Is Not a Media Company: "I Consider Us to Be a Technology Company"' (*CNBC*, 11 April 2018) https://www.cnbc.com/2018/04/11/mark-zuckerberg-facebook-is-a-technology-company-not-media-company.html.

created by their users and platforms provide the technology that enables the distribution of such content. On this ground, platforms have been shielded, for instance, from the application of public interest obligations imposed on traditional media, such as providing access to minimum levels of public programming, minimum levels of local content, and ensuring equal coverage of political candidates at a time of election.

However, as their centrality for public discourse came to grow, so did the criticism towards this overlooking of platforms' increasingly political and cultural role in controlling content that they broadcast to millions or even billions of users.[22] Among the criticisms are that platforms are not passive intermediaries, as they do not only moderate, but also recommend and curate content;[23] that the very use of the word "platform" is problematic because it downplays the fact that their services are not equally and meritocratically available to everyone, in particular if they do not take sufficient steps to prevent trolling and harassment;[24] that platforms can shirk responsibilities for their public footprint, and hide all of the labor necessary to produce and maintain their services;[25] and that platforms do not do enough to detect and remove illegal content of various kinds, such as for instance copyright-infringing material, sex trafficking and terrorist content. This dissatisfaction with platforms' efforts to tackle illegal content has grown incrementally over the last

[22] Philip Napoli and Robyn Caplan, 'Why Media Companies Insist They're not Media Companies, Why They're Wrong, and Why It Matters' (2017) 22 First Monday (online).

[23] Tarleton Gillespie, 'Platforms Are Not Intermediaries' (2018) 2 Georgetown Technology Law Review 198.

[24] Tarleton Gillespie, 'The Platform Metaphor Revisited' (*Digital Society Blog*, 24 August 2017) https://www.hiig.de/en/the-platform-metaphor-revisited/.

[25] Ibid.

few years, leading to a set of initiatives aimed to introduce new obligations,[26] some of which have made their way into legislation.[27]

Facebook has found itself front and center in this movement, particularly as a result of a series of media scandals that contributed both to the formation of strong public opinions and to the redefinition of the company's vision on these matters. For example, in his formal remarks to the US House Committee on Energy and Commerce in March 2018, Facebook's CEO Mark Zuckerberg recognized having failed to meet the company's responsibility in preventing its technology to be used for harm, specifically referring to the scandals involving the Cambridge Analytica data breach and the Russian interference in the US election.[28] In the following year, Zuckerberg made a similar statement about Facebook's responsibility to keep people safe and called for more active regulation, admitting that the company has too much power over speech, and announcing the creation of an independent body to appeal content moderation decisions,[29] what has come to be known as the Oversight Board. Facebook recently reinforced this call by

---

[26] In the EU see e.g. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Online Platforms and the Digital Single Market Opportunities and Challenges for Europe (COM(2016)288); Commission Recommendation of 1.3.2018 on measures to effectively tackle illegal content online (C(2018) 1177 final); Proposal for a Regulation on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC (COM(2020) 825 final 020/0361(COD).

[27] In the EU for example, see Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market; Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services; Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA.

[28] Dawn C. Chmielewski, 'Mark Zuckerberg Tells Congress: "I Started Facebook. I Run It. I'm Responsible For What Happens Here"' (*Deadline*, 09 April 2018) https://deadline.com/2018/04/mark-zuckerberg-tells-congress-i-started-facebook-run-it-im-responsible-1202361385/.

[29] Mark Zuckerberg, 'Online Post' (*Facebook*, 30 March 2019) https://www.facebook.com/4/posts/10107013839885441.

charting a way forward for Internet regulation[30] and launching a page entitled "It's time for updated Internet regulations," where it notes that tech companies need "standards that hold them accountable" and that it has been "a quarter-century" since comprehensive (US) Internet regulations were passed.[31]

MARK ZUCKERBERG: Chairman Grassley, Chairman Thune, Ranking Member Feinstein, Ranking Member Nelson and members of the committee, we face a number of important issues around privacy, safety and democracy. And you will rightfully have some hard questions for me to answer. Before I talk about the steps we're taking to address them, I want to talk about how we got here.

Facebook is an idealistic and optimistic company. For most of our existence, we focused on all of the good that connecting people can do. And, as Facebook has grown, people everywhere have gotten a powerful new tool for staying connected to the people they love, for making their voices heard and for building communities and businesses.

Just recently, we've seen the "Me Too" movement and the March for our Lives organized, at least in part, on Facebook. After Hurricane Harvey, people came together to raise more than $20 million for relief. And more than 70 million businesses — small business use Facebook to create jobs and grow.

But it's clear now that we didn't do enough to prevent these tools from being used for harm, as well. And that goes for fake news, for foreign interference in elections, and hate speech, as well as developers and data privacy.

ZUCKERBERG: We didn't take a broad enough view of our responsibility, and that was a big mistake. And it was my mistake. And I'm sorry. I started Facebook, I run it, and I'm responsible for what happens here.

Figure 1: Transcript from Zuckerberg's Senate Commerce and Judiciary committees testimony, April 10, 2018.

This growing recognition of responsibility is not unique to Facebook. Twitter, for example, famously described itself as "the free wing of the free speech party" to denote their liberal

[30] Monika Bickert, 'Charting a Way Forward Online Content Regulation' (*Facebook*, February 2020) https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf.

[31] Facebook, 'Internet Regulations - About Facebook' (*Facebook*, 2021) https://about.fb.com/regulations.

approach to content moderation:[32] they originally believed they were not in the business of deciding what is good or bad content, and that good speech is "the most effective antidote" to bad speech.[33] A few years in, however, the company's position changed as it became clear that they could only stand for freedom of expression if people feel safe to express themselves in the first place. Under this revised approach, Twitter began removing racism, extremism, and abuse, hate symbols and violent groups.[34] Commenting on this move, Twitter's former CEO Jack Dorsey referred to the need to balance free speech, safety and privacy[35], which map onto three of the five overarching values that Facebook recognized as informing its community standards (the other two being dignity and authenticity).[36]

Recent events illustrate how both Twitter and Facebook took seriously this more balanced approach to speech, and the repercussions of this approach on the tech industry more broadly. On January 7, 2021, following Donald Trump's use of Twitter to condone the actions of his supporters who rioted on Capitol Hill in an attempt to overturn his defeat at the US presidential election, Twitter first temporarily and then permanently suspended Trump's account, on grounds of risk of further incitement of violence.[37] On the same day, Facebook reached an similar decision (indefinite suspension), taking issue with Trump's use of the platform to incite violent insurrection

---

[32] Josh Halliday, 'Twitter's Tony Wang: "We Are the Free Speech Wing of The Free Speech Party"' *The Guardian* (22 March 2012).

[33] Shona Ghosh, 'Twitter Was Once a Bastion of Free Speech But Now Says It's "No Longer Possible to Stand up For All Speech"' (*BusinessInsider*, 19 December 2017) https://www.businessinsider.com/twitter-no-longer-possible-to-stand-up-for-all-speech-2017-12.

[34] Ibid.

[35] Nicholas Thompson, 'Jack Dorsey on Twitter's Role in Free Speech and Filter Bubbles' (*Wired*, 16 October 2010) https://www.wired.com/story/jack-dorsey-twitters-role-free-speech-filter-bubbles/.

[36] Monika Bickert, 'Updating the Values That Inform Our Community Standards' (*Facebook*, 12 September 2019) https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards/.

[37] Twitter, 'Permanent Suspension of @realDonaldTrump' (*Twitter*, 8 January 2021) https://blog.twitter.com/en_us/topics/company/2020/suspension.html.

against a democratically elected government.[38] In a matter of days, similar actions were taken by Google and Snapchat (respectively, suspending Trump's YouTube channel; and permanently suspending his account as of his last day in the office); and even Reddit (a notoriously liberal community-driven website, which leaves the bulk of decisions to volunteer moderators of so called "subreddits") decided to ban some pro-Trump forums.[39] This caused Trump supporters and right-wing groups to flock in mass to less well-known and openly "moderation-free" social media, such as Parler, Gab and Clapper, all of which were subsequently pressured from companies operating higher up in the technology stack (Amazon with its AWS servers, Apple and Google with their app stores)[40] who leveraged their position to marginalize platforms opting for "free speech maximalism."

[38] Guy Rosen and Monika Bickert, 'Our Response to the Violence in Washington (*Facebook*, 06 January 2021) https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/.

[39] Dylan Byers, 'How Facebook and Twitter decided to Take Down Trump's Accounts' (*NBC*, 14 January 2021) https://www.nbcnews.com/tech/tech-news/how-facebook-twitter-decided-take-down-trump-s-accounts-n1254317.

[40] Jerusalem Demsas, 'The Online Far Right is Angry, Exultant, and Ready for More' (*Vox*, 11 January 2021) https://www.vox.com/2021/1/9/22220716/antifa-capitol-storming-far-right-trump-biden-election-stop-the-steal-hawley-cruz; Makena Kelly, 'Clapper Permanently Bans QAnon-related Content' (*The Verge*, 11 February 2021) https://www.theverge.com/2021/2/11/22278480/clapper-tiktok-clone-bans-qanon-content-parler-deplatforming-capitol-riot; Copia Institute, 'Content Moderation Case Study: Decentralized Social Media Platform Mastodon Deals With An Influx Of Gab Users (2019)' (*Techdirt*, 3 March 2019) https://www.techdirt.com/articles/20210303/14474346357/content-moderation-case-study-decentralized-social-media-platform-mastodon-deals-with-influx-gab-users-2019.shtml.

Figure 2: Facebook's block of Trump's Facebook and Instagram accounts following the Capitol Hill incident on January 6, 2021 on the grounds of incitement to violence.

In just few days, it became apparent that an extremely libertarian attitude is no longer a viable proposition in the industry, thus marking a departure from the speech-protective policies that have tended to characterize major internet platforms, as inspired by the North American tradition.[41]

At the same time, it would be wrong to conclude that this industry shift has been well received by regulators around the world. In fact, Facebook's and Twitter's decisions in this de-platforming saga triggered a substantial amount of controversy and disagreement, including harsh criticism by heads of States such as former German Chancellor Angela Merkel, UK Prime Minister Boris Johnson and Mexico's President Andres Manuel Lopez Obrador.[42] Taking criticism one step

[41] Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech" (2018) 131 Harvard Law Review 1598.

[42] 'A "Bad Sign": World Leaders and Officials Blast Twitter Trump Ban' (*Al Jazeera*, 11 January 2021) https://www.aljazeera.com/news/2021/1/11/a-bad-sign-world-leaders-and-officials-blast-twitter-trump-ban.

further, the Polish government announced the tabling of a legislative bill that would make it illegal for tech companies to take similar actions (and including a "freedom of speech council" that would be able to order social networks to restore removed content).[43] This story vividly illustrates the palpable tension between State authorities and the new sovereigns in cyberspace, who set and enforce their own rules in virtual isolation from the "law of the land."[44] The growing tension between the law of the land and the law of the platform has also been manifest in recent episodes involving world leaders. For instance, in March 2020 Brazilian President Jair Bolsonaro had his posts removed by Twitter, Facebook and YouTube for including misinformation about the effectiveness of hydroxychloroquine as a treatment for Covid-19, which was held to violate social media platforms' rules against posting harmful content.[45] By the same token, Twitter removed a post about home-made treatment for Covid-19 by Venezuelan President Nicolás Maduro.[46]

---

[43] Adam Easton, 'Poland Proposes Social Media 'Free Speech' Law (*BBC*, 15 January 2021) https://www.bbc.co.uk/news/technology-55678502.

[44] Luca Belli, Pedro Augusto Francisco and Nicolo Zingales, 'Law of the Land or Law of the Platform?: Beware of the Privatisation of Regulation and Police' in Luca Belli and Nicolo Zingales (eds), *Platform Regulations: How Platforms Are Regulated And How They Regulate Us* (FGV Direito Rio, 2017).

[45] Kurt Wagner, 'Facebook, Twitter, YouTube Remove Posts From Bolsonaro' (*Bloomberg*, 30 March 2020) https://www.bloomberg.com/news/articles/2020-03-31/facebook-twitter-pull-misleading-posts-from-brazil-s-bolsonaro. Perhaps not coincidentally, in January 2021, some parliamentarians from Bolsonaro's party tabled a legislative proposal which would make it illegal for Internet platforms to remove content without prior judicial order. See: Requirement n° 211/2021 (Brazil) https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra;jsessionid=node018tjex8l568jtfierjkggkvgc8100119.node0?codteor=1964237&filename=Tramitacao-PL+213/2021.

[46] 'Coronavirus: World Leaders' Posts Deleted over Fake News' (*BBC*, 31 March 2021) https://www.bbc.co.uk/news/technology-52106321.

3. The entanglement of platform law and other fields of law[47]

The law of the platform is not developed in a vacuum: it is actively shaped by the interaction(s) of a plurality of actors.[48] Social media platforms' content moderation practices are influenced and affected by states' regulatory efforts through *content restrictions laws*, which define categories of content as illegal in particular domestic or regional contexts, and through *intermediary liability laws*, which set out the criteria under which intermediaries may be held liable for unlawful content generated by their users.[49]

International human rights law also influences and affects content moderation policies: by virtue of the UNGPs,[50] business enterprises, including social media companies, have a responsibility to respect human rights and to offer access to remedy when human rights are adversely impacted by their business activities. The operationalization of the UNGPs in the context of content moderation has been delineated comprehensively in the 2018 Report of the UN Special Rapporteur on Freedom of opinion and expression,[51] which is widely considered to be a landmark report as it translates the application of human rights standards to content moderation specifically.

The exercise of online freedom of expression is thus increasingly regulated by different layers, which in some instances appear to be aligned, but in others stand in sharp contrast. Disentangling and navigating the interactions between these layers is not an easy task. Content restrictions laws may in fact originate at the international, regional or domestic level, and, for instance, international

---

[47] To further explore the notion of "legal entanglements" see N. Krisch (ed) *Entangled Legalities Beyond the State* (Cambridge University Press, 2021).

[48] Barrie Sander, 'Freedom of Expression in the Age of Online Platforms: Operationalizing a Human Rights-Based Approach to Content Moderation' (2020) 43 Fordham International Law Journal 939.

[49] Ibid 7.

[50] Guiding Principles on Business and Human Rights.

[51] David Kaye, 'Report of The Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (2018) A/HRC/38/35.

or regional regimes may impose an obligation upon states to criminalise specific categories of content.

An example of content restrictions laws at the regional level is provided by Council of Europe's Convention on Cybercrime,[52] the first international treaty addressing crimes committed via the Internet and other computer networks, and its Additional Protocol concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems.[53] Both instruments were drafted as a response to the emergence of international communication networks like the Internet, which were deemed to provide a new platform to "support racism and xenophobia and […] disseminate easily and widely expressions containing such ideas."[54] Similarly, international human rights law defines categories such as child pornography, direct and public incitement to genocide, advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence and incitement to terrorism as exceptional types of expression that states are required to prohibit.[55] When domestic legal regimes and the law of the platform incorporate these categories, we observe an overall alignment between these normative layers. But domestic legal regimes may nonetheless criminalize or otherwise restrict categories of content irrespective of the incompatibility of these restrictions with human rights standards. For instance, the Thai lèse-majesté laws, which have been heavily criticized by United Nations human rights experts,[56] prescribe that anyone who "defames, insults or threatens the king, the queen, the heir-apparent or the regent" will be punished with a jail term between 3 and 15 years. As the law of the land and international human rights law are in this case at odds,

---

[52] Council of Europe Convention on Cybercrime of the Council of Europe (2001).

[53] Council of Europe Additional Protocol to the Convention on Cybercrime, Concerning the Criminalisation of Acts of a Racist and Xenophobic Nature Committed Through Computer Systems' (2003).

[54] Ibid 1.

[55] Frank La Rue, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (2011) A/HRC/17/27.

[56] United Nations, 'Thailand: UN Experts Alarmed by Rise in Use of Lèse-Majesté Laws" (*United Nations*, 8 February 2021) https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=26727.

tensions may arise between the law of the platform and the law of the land or international human rights standards.

With the creation of the Oversight Board, the entanglement(s) between the law of the platform and other legal regimes become even more evident. The Oversight Board is tasked with reviewing content decisions made by Facebook (on the basis of their content moderation policies) that have been appealed by users.[57] When reviewing content decisions, the Oversight Board applies the law of the platform (Facebook's content policies and values), but it is also required to "pay [...] attention to the impact of removing content in light of human rights norms protecting free expression."[58] The first set of the Board's decisions reveals a strong entanglement between the law of the platform and international human rights law: in each decision, the Board has provided a thorough analysis of Facebook's content decisions not only in light the company's policies and values, but also in light of international human rights law, taking into consideration a wide range of instruments (treaty provisions and authoritative guidance of the UN's human rights mechanisms). Since the Board also provides Facebook with policy guidance in light of the decisions taken, international human rights law might now inform the development of Facebook's content policies more substantively. At the same time, an assessment of the interactions between domestic law(s) and the law of the platform is currently obstructed by the fact that the Board cannot review content whose reinstatement would (1) violate criminal law in the national jurisdiction(s) concerned or (2) result in "adverse governmental action" against Facebook because of its unlawfulness in the national jurisdiction(s) concerned.[59]

The growing power of online platforms, including Facebook, and the concomitant concerns around how they use their power has resulted in concrete enforcement activity and calls for even more in the future. Regulators have started growing uneasy about the status of online platforms as "infrastructural intermediaries"[60] that serve as "custodians to the massive, heterogeneous, and

---

[57] Facebook, 'Establishing Structure and Governance for an Independent Oversight Board' (*Facebook*, 8 February 2021) https://newsroom.fb.com/news/2019/09/oversight-board-structure/.

[58] Oversight Board Charter 2019, Article 2§2.

[59] Oversight Board Bylaws 2020, Article 2§1.2.2.

[60] Paul Langley and Andrew Leyshon, 'Platform capitalism: The Intermediation and Capitalization of Digital Economic Circulation' (2016) 3(1) Finance and Society (online).

contested public realm they have brought into being"[61] and at the same time as amplifiers of harmful speech at scales that were unheard of a few years ago, even after the Internet ear had well arrived.[62] Interestingly, the efforts to rein in online platforms has spanned various fields of law—indicative of their pervasive presence.

4. The evolving framework for intermediary responsibility

To understand how the discourse on the responsibility of social media platforms has evolved over the past 15 years, it is essential to appreciate the legal framework that permitted the emergence and growth of these platforms. In this Section, we describe the regime in place in the United States and Europe (with occasional references to other regimes), who were the early adopters of a set of specific rules regulating the responsibility of intermediaries, defined here broadly as the entities that provide services that enable communication between individuals. Prior to examining that specific set of rules, however, it is worth noting that liability for third party content (also known as "secondary liability") is traditionally established under tort law on the basis of two different theories: contributory liability and vicarious liability. The former presupposes that the secondary infringer has knowledge of the infringing activity and makes a material contribution to it; whereas the latter does not require knowledge, it simply stands upon a relationship between the two joint tortfeasors, where one is in control of the other's activity and derives a financial benefit from it. Although these theories provide foundational principles for the development of secondary liability, this is simply insufficient to guide an intermediary's conduct in the variety of complex situations that it faces in the current technological age, where intermediation is embedded in our every single action online. Indeed, this is both the point of departure and the

[61] Tarleton Gillespie, *Custodians of The Internet: Platforms, Content Moderation, and The Hidden Decisions That Shape Social Media* (Yale University Press 2018) 211.

[62] Michael Krzyżanowski, 'Normalization and The Discursive Construction of "New" Norms and "New" Normality: Discourse in The Paradoxes of Populism and Neoliberalism' (2020) 30 Social Semiotics (online).

aspiration behind legislation specifically focused on the responsibility of Internet intermediaries: the need to provide legal certainty and stimulate the growth of Internet services.

a) Communications Decency Act (CDA) §230 and its erosion

The first legislation that specifically addressed secondary liability on the Internet was Section 230 of the US Communication Decency Act (CDA). Specifically, this legislation introduced in 1996 by Senators Chris Cox and Ron Wyden focused on the broader concept of "interactive computer service" defined as "any information service, system, or access software provider that provides or enables computer access by multiple users to a computer server, including specifically a service or system that provides access to the Internet and such systems operated or services offered by libraries or educational institutions."[63] The statute explicitly ruled out the qualification of providers or users of an interactive computer service as "publishers" of any information provided by another information content provider (§230(c)(1)), and provided them with complete immunity from civil liability for any action voluntarily taken in good faith (so-called "good samaritan" behavior) to restrict or enable restriction of access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected (§230(c)(2)).

The motivation and rationale behind the introduction of Section 230 was one of direct response to litigation involving websites hosting online news forums, back then a very popular channel for interaction among internauts. Three cases, in particular, had made evident the threat that secondary liability could pose to the development of websites and other novel Internet applications, as much as to the maintenance of a civil online public sphere that allows citizens to

[63]See Section 230(f)(2) and (3). Furthermore, Section 230(f)(4) clarifies that "access software provider" refers to a provider of software or enabling tools that do any one or more of the following: Filter, screen, allow, or disallow content; Pick, choose, analyze, or digest content; or transmit, receive, display, forward, cache, search, subset, organize, reorganize, or translate content.

enjoy their freedom of expression:[64] *Cobby v Computerserve*,[65] *Playboy v. Frena,*[66] and *Stratton Oakmont Inc v. Prodigy.*[67] Of these cases, the first (*Cobby*) ruled against the extension of publisher liability to websites hosting third party content, while the other two suggested the existence of a moderation paradox: if websites took the decision to moderate the content they hosted, they ran into the risk of being liable for any of the infringing content that they let slip through, as they were deemed to have editorial control over it. Thus, the passing of Section 230 which directly overruled this stance was perceived by courts as a strong signal in favor of the protection of websites for hosting content, with the aim to "maintain the robust nature of the Internet" in the face of "the threat that tort-based lawsuits pose to freedom of speech in the new and burgeoning Internet medium."[68] Not coincidentally, Section 230(c)(1) has been called "the twenty-six words that created the Internet."[69]

This notwithstanding, litigation throughout the last 25 years over Section 230 exposed the potentially very serious impact on freedom of speech following from a broad interpretation of the statute: in *Noah v AOL*,[70] for instance, Section 230 prevented the plaintiff from making a chat-room accountable for the anti-Islamic slurs that it permitted, thereby making the environment inhospitable to Islamic users. While on this occasion the Court of Appeal of the 4th Circuit saw Section 230 as an impenetrable shield against these types of claims, the Court of Appeal for the 9th Circuit adopted a diametrically different posture in *FHC v Roommates* with regard to the facilitation of discrimination: it reasoned that a website for classified ads is a publisher of speech that it contributed to create, in particular through a drop-down menu that permitted users to rely

---

[64] Matt Reynolds, 'The Strange Story of Section 230, The Obscure Law That Created Our Flawed, Broken Internet' (*Wired UK*, March 24 2019) https://www.wired.co.uk/article/Section-230-communications-decency-act.

[65] *Cubby, Inc. v. CompuServe* Inc, 776 F. Supp. 135 (S.D.N.Y., 1991).

[66] *Playboy Enterprises, Inc. v. Frena*, 839 F.Supp. 1552 (1993).

[67] *Stratton Oakmont, Inc. v. Prodigy Services Co.*, WL 323710 (N.Y. Sup. Ct., 1995).

[68] *Zeran v. America Online, Inc.*, 129 F.3d 327 (4th Cir. 1997).

[69] Jeff Kossef, *The Twenty-Six Words That Created the Internet* (Cornell University Press, 2019).

[70] *Noah v. AOL Time Warner, Inc.*, 261 F. Supp. 2d 532 (Eastern District of Virginia, 2003).

on discriminatory criteria to search for housemates.[71] More recent cases follow this path to further curtail the scope of immunity: for instance, in *Jones v Dirty*, the district court held to the standards of publisher liability a website that encouraged the posting of "gossip" that proved to be defamatory. However, the 6th Circuit court of appeal reversed, clarifying that the governing test is not one of "encouragement", but one of "material contribution."[72] In *Oberdorf v. Amazon.com Inc*,[73] the Court of Appeal for the 3rd Circuit clarified that "materially contribute" may also involve a failure to act: it ruled that, notwithstanding Section 230, Amazon can be held liable for the sale of defective products on its platforms because it is uniquely positioned to receive reports of defectiveness. This ruling built on a 9th Circuit decision in *Model Mayhem*, where Section 230 was held inapplicable to a website that provided a service of matching for models with modeling jobs, finding that it had a duty to warn models about individuals who were known to be using the website to find women to sexually assault.[74]

Concerns with the breadth of the immunity provided by Section 230 prompted numerous attempts of reform, especially over the last 5 years. Of these, so far only one has been successful: on April 11, 2018, President Donald Trump signed into law the Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA) and the Stop Enabling Sex Traffickers Act (SESTA) — a combination of bills passed by the House and Senate. This law increased the areas exempted from Section 230 immunity, which originally include federal criminal liability (§230(e)(1)), electronic privacy violations (§230(e)(4)) and intellectual property claims (§230(e)(2), to cover also federal and state claims related to sex trafficking.[75].

---

[71] *Fair Housing Council of San Fernando Valley v. Roommates.com, LLC* 521 F.3d 1157 (9th Cir. 2008).

[72] *Jones v. Dirty World Entm't*, 755 F.3d 398 (6th Cir. 2014).

[73] *Oberdorf v. Amazon.com Inc*., 936 F.3d 182 (3d Cir. 2019).

[74] Gus Hurwitz, 'The Third Circuit's Oberdorf v. Amazon Opinion Offers a Good Approach to Reining in The Worst Abuses of Section 230' *(Truth on the Market*, 15 July 2019) https://truthonthemarket.com/2019/07/15/the-third-circuits-oberdorf-v-amazon-opinion-offers-a-good-approach-to-reining-in-the-worst-abuses-of-section-230/.

[75] In particular, Section 4 of the law provides that Section 230 does not limit: (1) a federal civil claim for conduct that constitutes sex trafficking, (2) a federal criminal charge for conduct that constitutes sex

The number of bills introduced in the last couple of years that would further erode Section 230 immunity is a telling sign of the perceived need to rebalance the social contract with digital platforms, and of the flurry of legislative activity in this space. These include:

- the Ending Support for Internet Censorship Act,[76] which in exchange for continued Section 230 protection would require entities with over 30 million active monthly users in the United States, over 300 million worldwide active monthly users, or more than $500 million in global annual revenue to obtain a certification with the Federal Trade Commission, having proven that they do not moderate information provided by other information content providers in a manner that is biased against a political party, political candidate, or political viewpoint.
- the Break Up Big Tech Act,[77] which would eliminate Section 230 protections for online services that (a) sell advertisements displayed to users based on their personal traits and behavior without opt-in; (b) place or facilitate the placing in commerce of certain items; (c) collect data for commercial purposes other than receiving from users of such service direct payment for the use of such service; or (d) use a design or product that addicts, or whose purpose is to addict, users to such service. Moreover, owners or operators of a social media service that display user-generated content in an order other than chronological order, delay the display of such content relative to other content, or otherwise hinder the display of such content relative to other content, if for a reason other than to execute a user request or to restrict access to or availability of material described in (a) would also be treated as publishers or speakers of such content.

trafficking, or (3) a state criminal charge for conduct that promotes or facilitates prostitution in violation of this bill.

[76] Ending Support for Internet Censorship Act of 2019, S.1914, 116th Cong.

[77] Break Up Big Tech Act of 2020, H.R.8922, 116th Cong.

- the Eliminating Abusive and Rampant Neglect of Interactive Technologies (EARN IT) Act,[78] which would introduce an exception to Section 230 with respect to claims alleging violations of child sexual exploitation laws.
- the Limiting Section 230 Immunity to Good Samaritans Act,[79] which would clarify the meaning of "good faith" and strip away Section 230 immunity for all websites, online applications, or mobile applications with over 30 million monthly U.S. users and over US$1.5 billion in global revenues which serve to distribute information provided by another information content provider (so called "edge providers") in case of violation of such duty, including in case of "selective enforcement" of terms of service.
- the Behavioral Advertising Decisions Are Downgrading Services Act,[80] which would remove Section 230 protections for large service providers (30 million users in the U.S. or 300 million globally and with more than US$1.5 billion in annual revenue) to the extent they use behavioral advertising.
- the Online Freedom and Viewpoint Diversity Act,[81] which would on the one hand replace the term "objectionable" with more specific categories, namely "promoting self-harm, promoting terrorism, or unlawful" in relation to the types of moderations that would not give rise to editorial responsibility; and on the other hand, would attribute such responsibility where a person or entity editorializes or affirmatively and substantively modifies the content of another person or entity, with the exception of mere formatting changes.
- the Safeguarding Against Fraud, Exploitation, Threats, Extremism, and Consumer Harms Act (SAFE TECH act),[82] which would remove the liability protection against injunctive relief (as opposed to damages actions), limit such protections to the carrying of third-party "information" (as opposed to "speech"), and make further carveouts for

[78] Eliminating Abusive and Rampant Neglect of Interactive Technologies Act of 2020, S. 3398, 116th Cong.

[79] Limiting Section 230 Immunity to Good Samaritans Act, S.3983, 116th Cong.

[80] Behavioral Advertising Decisions Are Downgrading Services Act 2020, S. 4337, 116th Cong.

[81] Online Freedom and Viewpoint Diversity Act 2020, S.4534, 116th Cong.

[82] Safe Tech Act of 2021, S. 299, 117th Cong.

claims arising under civil rights laws, antitrust laws, cyberstalking laws, human rights laws or civil actions regarding a wrongful death, as well as cases in which the provider is receiving a payment to make the information available.

Although the above-mentioned proposals come from across the entire political spectrum, it should not be concluded that this is merely a political debate. As testament of that, the Department of Justice completed a review of Section 230 protections and issued four key recommendations to Congress in June 2020:[83]

- Limiting protections only for "responsible" online platforms, to the exclusion of: those that facilitate or solicit third party content or activity that violates federal criminal law; those that have specific knowledge of infringing content or activity; and those that host particularly egregious content such as child exploitation and sexual abuse, terrorism and cyber- stalking.
- Removing protections from civil lawsuits brought by the federal government.
- Clarifying that federal antitrust claims are not covered by Section 230 immunity.
- Promoting Open Discourse and Greater Transparency by clarifying terms such as "otherwise objectionable" and "good faith", and explicitly stating that a provider does not lose immunity simply because it removes content pursuant to Section 230(c)(2) or consistent with its terms of service.

It should be clear by now that the Department of Justice is substantially in agreement with the push for reform, with its substantive suggestions revolving on some of the recurring themes of the reform proposals. It is also interesting to see that the idea of narrowing the scope of Section 230 immunity has found support at the Supreme Court: in a recent opinion, Justice Thomas persuasively made the argument that Section 230 (a) should extend only to publishers, not distributors; (b) should not extend to online platforms' selection and editing of third-party content; (c) should be interpreted narrowly so as not to swallow up Section 230(c)(2)'s "Good Samaritan"

[83] Department of Justice Review of Section 230 of The Communications Decency Act of 1996 (*United States Department of Justice*, 17 June 2020) https://www.justice.gov/archives/ag/department-justice-s-review-section-230-communications-decency-act-1996.

immunity for online platforms' good faith removal of objectionable content; and (d) should not be interpreted to preclude traditional product-defect claims.[84]

b) Digital Millennium Copyright Act (DMCA) § 512 and its misuse

Two years after the introduction of CDA § 230, US Congress passed another landmark legislation providing liability limitations for online intermediaries, in the specific context of copyright law. The Digital Millennium Copyright Act, and in particular Section 512, introduced detailed rules for the limitation of liability of four types of intermediaries in the copyright context. These rules have been highly influential for the development of intermediary liability protections in various jurisdictions around the world, both within and outside the copyright context.

While the DCMA identifies different types of intermediation which are granted immunity from liability (a "safe harbour"), only one is relevant in the particular context of Facebook's social networking services: the so-called "hosting", which refers to storage occurring "at the direction of a user of material that resides on a system or network controlled or operated by or for the service provider."[85] For instance, this would include cloud computing services or simple email storage, social media, etc. A provider of these services benefits from safe harbour only upon fulfilling the following conditions:

- Does not have actual knowledge of the infringing nature of the material, and is not aware of facts or circumstances from which infringing activity is apparent; or upon obtaining such knowledge or awareness, acts expeditiously to remove, or disable access to, the material; and
- Does not receive a financial benefit directly attributable to the infringing activity, in a case in which the service provider has the right and ability to control such activity; and upon notification of claimed infringement, responds expeditiously to remove, or disable access; and

[84] *Malwarebytes, Inc., Petitioner v. Enigma Software Group USA, LLC*, 946 F.3d 1040 (9th Cir. 2019).

[85] 17 U.S.C. § 512(c).

- Has a designated agent for the notification of claims of infringements and follows the special procedure of notice and take-down indicated by Section 512(g).

In turn, the procedure stipulated in Section 512 (g) prescribes that reasonable steps must be taken promptly to notify the subscriber that the provider has removed or disabled access to the material; that upon receipt of a counter notification, the provider promptly provides the person who provided the notification with a copy of the counter-notification, and informs that person that it will replace the removed material or cease disabling access to it in 10 business days; and that after no less than 10 and no more than14 business days following a counter-notification, it must replace the removed material and cease disabling access to it unless its designated agent first receives notice from the person who submitted the notification that such person has filed an action seeking a court order to restrain the subscriber from engaging in infringing activity relating to the material on the service provider's system or network.

However, paragraph (1) of Section 512 (g) also provides that a service provider shall not be liable for any claim based on the service provider's good faith disabling of access to, or removal of, material or activity claimed to be infringing, or based on facts or circumstances from which infringing activity is apparent, *regardless of whether the material or activity is ultimately determined to be infringing* (emphasis added). By contrast, there is no equivalent provision for the failure to remove the material or activity, regardless of whether the counternotice process has been triggered. As a result, an implication of Section 512 (g) is that hosting providers, when in doubt about the legality of a particular content or activity, will have a clear incentive to remove or disable access to it in order to escape possible liability.[86] The chilling effect on freedom of expression is compounded by the fact that empirical research also documents a large number of questionable legal claims in those original notices.[87] However, this effect might have been mitigated following the ruling in *Lenz v. Universal*, where the Court of Appeals for the 9th Circuit

[86] Jennifer M. Urban, Joe Karaganis, and Briana Schofield, 'Notice and Takedown in Everyday Practice' (2016) UC Berkeley Public Law Research Paper No. 2755628 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2755628.

[87] Ibid.

held copyright holders have a duty to consider "in good faith and prior to sending a takedown notification" whether the allegedly infringing content is protected by fair use.[88]

Another controversial issue about section 512 is the extent to which hosting providers may be called to respond for the infringing content available on their sites when they have no specific knowledge of the infringement. The statute talks about awareness of "facts or circumstances from which infringing activity is apparent", also known as "red flag" knowledge, and this has been interpreted by the 2nd Circuit Court of Appeal in *Viacom International v. Youtube*[89] to refer to situations in which the provider was subjectively aware of facts that would have made a specific infringement "objectively" obvious to a reasonable person. The most unpredictable part of this test for a provider revolves around what circumstances are deemed sufficient to establish such obviousness: for instance, the 9th Circuit Court of Appeal in *Columbia Pictures Industries, Inc. v. Fung* deemed it sufficient that the provider had been "actively encouraging infringement, by urging his users to both upload and download particular copyrighted works, providing assistance to those seeking to watch copyrighted films, and helping his users burn copyrighted material onto DVDs.[90]" Similarly, the 2nd Circuit Court of Appeal in *EMI Christian Music* found the threshold met when a site was conceived of and designed to facilitate infringement[91], while also making clear that in such cases there is no general duty to monitory, but simply a time-limited, targeted duty—even if encompassing a large number of songs[92]. Another type of situation that would deprive the provider of the safe harbour is the so-called "wilful blindness", in other words when a provider blinded itself to possible exposure to infringing activity by its users despite awareness of the high probability of the fact in dispute[93]. Whereas this doctrine has been interpreted narrowly, requiring this probability to be of a *specific* infringement, the US Copyright Office has recently criticized this interpretation

---

[88] *Stephanie Lennz v. Universal Music*, 815 F.3d 1145 (9th Cir. 2015).

[89] *Viacom International v. Youtube*, 676 F.3d at 34 (2d Cir. 2012).

[90] *Columbia Pictures Industries, Inc. v. Fung*, 710 F.3d (9th Cir. 2013) at 1043.

[91] *EMI Christian Music Grp., Inc. v. MP3tunes*, LLC, 844 F.3d 79 (2d Cir., 2016) at 93.

[92] Ibid.

[93] *Viacom* (n 89) 35.

in a report on Section 512 as being too narrow[94]. However, it also warned that Congress must strike a balance between increasing the effectiveness of copyright protection and imposing an appropriate burden on hosting providers, as broadening the application of this doctrine may result in a moderation paradox analogous to the one that CDA Section 230 aimed to avoid: to the extent that a provider moderates content on its site, it may be deemed to have knowledge about the infringing content that it makes available.

c) E-commerce Directive

Rules on the liability of Internet intermediaries in the European Union find a common root in the 2000 E-commerce Directive, which had as overarching goals the development of e-commerce, the achievement of a balance between conflicting fundamental rights and the sharing of responsibility between all the private actors of the ecosystem in ensuring the minimization of illegal material and a good cooperation with public authorities.

In drafting the Directive, legislators were inspired by the DMCA, but decided to replicate only a subset of the rules contained therein, and to extend its application beyond the realm of copyright infringements. Thus, Articles 12-15 contain the core provisions for the regime of liability of "information society service providers" in Europe. An "information society service" is defined as "any service normally provided for remuneration,[95] at a distance, by means of electronic equipment for the processing (including digital compression) and storage of data, and at the individual request of a recipient of a service.[96] However, three specific categories of intermediation are defined in Articles 12-14, including "conduit," "caching," and "hosting". As in the DCMA

[94] US Copyright Office, Section 512 Report (27 May 2020), available at https://www.copyright.gov/policy/section512/section-512-full-report.pdf, pp 127-128.

[95] Recital 19 clarifies that this is not the case, for example, for public education and governmental services.

[96] See Article 2 (a) of the European E-Commerce Directive 2000/31, referring to the definition in art. 1(2) of Directive 98/34, as amended by Directive 98/48.

context, we will focus here on "hosting", which is defined by Article 14 as "the storage of information provided at the request of a recipient of the service." The safe harbour is based on the following conditions:

- The provider does not have actual knowledge of illegal (either civil or criminal) activity or information, nor (as regard claims for damages) has awareness of facts and circumstances from which such illegality is apparent;
- Upon obtaining such knowledge or awareness, the provider acts expeditiously to remove or to disable access to the information;
- The provider has no authority or control over the recipient.

This safe harbour is incomplete, however, as it does not specify what counts as "actual knowledge." This has enabled EU member states to adopt different approaches in the implementation of the Directive, such as requiring a formal notification by the competent administrative authorities (Spain), the fulfilment of a notice and take-down procedure (Finland), or leaving the determination to national courts on a case by case basis (Germany and Austria).[97]

On top of that, a significant degree of uncertainty with regard to the intermediary's imputed knowledge was generated by the European Court of Justice's ruling in Google France, where the Court answered a preliminary reference on Google's possible liability for use of brand-related keyword by a brand's competitor by affirming that the safe harbour is linked to recital 42 of the Directive, which holds that the activity of the information society service provider must be "of a mere technical, automatic and passive nature," implying that that service provider "has neither knowledge of nor control over the information which is transmitted or stored." That recital was actually conceived of with Article 12 (conduit) providers in mind, and its extension to Article 14 (hosting) providers would imply a certain level of neutrality on the part of a host. The Court appears to have departed from that requisite in L'Oréal v eBay, suggesting that the reference criterion is not neutrality, but whether a "diligent economic operator" should have realized, based on its awareness of certain facts or circumstances, that the offers for sale in question were

[97] See Patrick Van Eecke, Maarten Truyens, 'EU Study on the Legal Analysis of a Single Market for the Information Society' (2014), 231-232 https://op.europa.eu/en/publication-detail/-/publication/a856513e-ddd9-45e2-b3f1-6c9a0ea6c722.

unlawful.[98] On this basis, it determined that the defendant who had promoted or optimized the presentation of the offers for sale in question could be subject to injunctions entailing the "prevention of future infringements of the same kind," (e.g., suspension of accounts or measures facilitating the identification of infringers operating in the course of trade); however, such injunctions must strike fair balance with freedom to conduct business of intermediary and with rights of privacy, data protection and freedom of expression of the infringer.

The E-Commerce Directive circumscribes the boundaries of its safe harbours in two ways. One limit is negative, in the sense that it prohibits member states from regulating inconsistently with the EU framework: in particular (and similarly to the DMCA), imposing general obligation on providers of services covered by Articles 12, 13 and 14 to monitor the information which they transmit or store, or to actively to seek facts or circumstances indicating illegal activity. Much debate has unfolded on this particular concept, but the key takeaway seems to be that only monitoring of specific content (in respect of both the protected subject matter and potential infringers) can be imposed, and not also monitoring for specific content (which would require screening content in its entirety).[99]

The other limit is positive, clarifying that member states may establish obligations for information society service providers promptly to inform the competent public authorities of alleged illegal activities undertaken or information provided by recipients of their service, or obligations to communicate to the competent authorities, at their request, information enabling the identification of recipients of their service with whom they have storage agreements. This is to be added to the general caveat made by recital 48 that the Directive is without prejudice to the possibility for Member States of requiring service providers, who host information provided by recipients of their service, to apply duties of care which can reasonably be expected from them, and which are specified by national law in order to detect and prevent certain types of illegal activities. While the exact limits of these duties of care have yet to be tested in court, it has been

[98] Case C-324/09 *L'Oréal SA and Others v. eBay International AG and Others* (2009) ECR I-6011.

[99] Martin Senftleben and Christina Angelopoulos, 'The Odyssey of the Prohibition on General Monitoring Obligations on the Way to the Digital Services Act: Between Article 15 of the E-Commerce Directive and Article 17 of the Directive on Copyright in the Digital Single Market' (2020) Center for Intellectual Property & Information Law Working Paper https://ssrn.com/abstract=3717022.

suggested that they would imply public rather than private law duties, as a different interpretation would defeat the purpose of the liability limitations; and similarly, that duties of care must concern obligations that are not explicitly exempted by the safe harbour, such as duties of information or mandatory dispute resolution procedures.[100]

Even the duties of information, however, must strike a fair balance between the various fundamental rights involved, as clarified by the European Court of Justice in ProMusica regarding any possible duty on internet access providers to retain and communicate the personal data of subscribers in the context of civil proceedings.[101] By extension, the same balancing applies to any obligations imposed to information society service providers by courts or administrative orders in relation to an identified infringement,[102] including preventing the availability of "equivalent content"[103] and to the imposition of filtering to prevent further infringement from already infringing users.[104]

---

[100] European Commission, 'Hosting Intermediary Services and Illegal Content Online An Analysis of the Scope of Article 14 ECD in Light of Developments in the Online Service Landscape: Final Report' (2018) https://op.europa.eu/en/publication-detail/-/publication/7779caca-2537-11e9-8d04-01aa75ed71a1/language-en.

[101] Case C-275/06 *Productores de Música de España (Promusicae) v Telefónica de España SAU* (2008) ECR I-00271.

[102] Articles 12(3); 13(2); 14(3) and 18 of Directive 2000/31/EC of the European Parliament and of the Council on certain legal aspects of information society services, in particular electronic commerce in the Internal Market.

[103] Case C-18/18, *Eva Glawischnig-Piesczek v Facebook*, ECLI:EU:C:2019:821, where the CJEU held that such obligations are not precluded as long as " the monitoring of and search for the information concerned by such an injunction are limited to information conveying a message the content of which remains essentially unchanged compared with the content which gave rise to the finding of illegality and containing the elements specified in the injunction, and provided that the differences in the wording of that equivalent content, compared with the wording characterising the information which was previously declared to be illegal, are not such as to require the host provider to carry out an independent assessment of that content".

[104] Case C-70/10 *Scarlet Extended SA v. SABAM* (2011) ECLI:EU:C:2011:771.

d) New rules of responsibility

For several years, the carefully designed (and further clarified) framework above proved adequate to serve the needs of the information society. In the last few years, however, pressure mounted to request online platforms to play a more active role in the detection and removal of illegal content. This was most evident in the EU's Communication on Online Platform[105] and its proposals for a Directive on copyright in the Digital Single Market,[106] and for an updated Audiovisual Media Services Directive in 2016;[107] the Communication on Tackling Illegal Content

[105] EU Commission, Communication: Online Platforms and the Digital Single Market Opportunities and Challenges for Europe, Brussels, COM(2016) 288/2. The Communication identified four guiding principles: (1) the creation of a level playing field for comparable digital services, (2) the responsible behaviour of online platforms to protect core values, (3) the transparency and fairness for maintaining user trust and safeguarding innovation and (4) open and non-discriminatory markets in a data-driven economy.

[106] Directive of the European Parliament and of the Council on copyright in the Digital Single Market (2016), COM/2016/0593.

[107] Directive of the European Parliament and of the Council amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services in view of changing market realities (2016), COM/2016/0287 final - 2016/0151.

Online[108] and Directive 2017/54 on combating terrorism,[109] in 2017; and the proposed Regulation on preventing the dissemination of terrorist content online, in 2018.[110]

In parallel to this, the European Court of Justice did not sit idly by, and contributed to redefine (in particular in copyright law) the responsibility of digital platforms. The ruling that moved the needle towards new grounds in 2017 was *Stichting Brein v Ziggo*[111], a preliminary reference procedure where the European Court of Justice was asked to clarify whether the operator of a platform that makes available to the public third-party uploaded copyrighted content and provides functions such as indexing, categorization, deletion and filtering of content may be liable for copyright infringement jointly with users of that platform. The Court responded interpreting the Information Society Directive (2001/29) broadly to include any "indispensable intervention" in the concept of "communication to the public" of Article 3, according to which: "Member States shall provide authors with the exclusive right to authorize or prohibit any communication to the public of their works, by wire or wireless means, including the *making available* to the public of their works in such a way that members of the public may access them from a place and at a time individually chosen by them." Particularly, the Court found that a platform operator makes this intervention with full knowledge of the consequences of its conduct, when it provides access to protected works, by indexing on that platform torrent files which allow users of the platform to locate those works and to share them within the context of a peer-to-peer network.[112]

---

[108] Communication from the Commission to the European Parliament, the Council, The European Economic and Social Committee and the Committee of the Regions Tackling Illegal Content Online Towards an enhanced responsibility of online platforms (2017), COM(2017)555. See also Commission's follow- up recommendation, 1.3.2018 C(2018) 1177 final.

[109] Official Journal of the European Union, L 88 (2017), 6-21.

[110] Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online A contribution from the European Commission to the Leaders' (2018) COM/2018/640 final.

[111] Case C-610/15 *Stichting Brein v Ziggo BV and XS4All Internet BV* (2017) ECLI:EU:C:2017:456.

[112] Ibid para 36.

In a more recent ruling,[113] the Court nuanced the important point that a video-sharing or a file-hosting and -sharing platform does not make a "communication to the public" of its user-generated content unless it contributes, beyond merely making that platform available, to giving access to such content to the public in breach of copyright: it clarified that this is the case, *inter alia*, where that operator has specific knowledge that protected content is available illegally on its platform and refrains from expeditiously deleting it or blocking access to it; or where that operator, despite the fact that it knows or ought to know, in a general sense, that users of its platform are making protected content available to the public illegally via its platform, refrains from putting in place the appropriate technological measures that can be expected from a reasonably diligent operator in its situation in order to counter credibly and effectively copyright infringements on that platform; or finally, where that operator participates in selecting protected content illegally communicated to the public, provides tools on its platform specifically intended for the illegal sharing of such content or knowingly promotes such sharing, which may be attested by the fact that that operator has adopted a financial model that encourages users of its platform illegally to communicate protected content to the public via that platform.[114]

Interestingly, the same interpretation of "communication to the public" for platforms going beyond the mere provision of physical facilities had been suggested in 2016 by the European Commission in its proposed Directive on copyright for certain types of information society service providers, holding that "[w]here information society service providers store and provide access to the public to copyright protected works or other subject-matter uploaded by their users, thereby going beyond the mere provision of physical facilities and performing an act of communication to the public, they are obliged to conclude licensing agreements with right-holders, unless they are eligible for the liability exemption provided in Article 14 of Directive 2000/31/EC."[115] It also required such entities to take appropriate and proportionate measures to ensure the protection of works or

---

[113] Joined Cases C-682/18 and C-683/18 *Frank Peterson v Google LLC and Others and Elsevier Inc.v Cyando AG* (2021) ECLI:EU:C:2021:503.

[114] Ibid para 102.

[115] Amendments adopted by the European Parliament on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market (2018), Recital 38.

other subject-matter, such as implementing effective technologies,[116] even where they would be eligible for the hosting safe harbour. This was the first explicit admission that EU law can impose duties that are antithetical to a passive position, without in doing so violating the principles established in the E-commerce Directive (and thus implicitly overruling *Google France*). The proposal was not immune from criticism, with, among others, several academics calling Article 13 an imposition of "general monitoring" obligations, in plain contrast with the prohibition for such obligations established in Article 15 of the E-commerce Directive.[117]

The legislation that was eventually produced (Directive 2019/790), after an intense discussion with industry and experts, revolved around the same concept of communication to the public with some significant differences in its operation and limits. Most importantly, for our purposes, the new version of Article 13 (Article 17) explicitly recognized that "when an online content-sharing service provider performs an act of communication to the public or an act of making available to the public under the conditions laid down in this Directive, *the limitation of liability established in Article 14(1) of Directive 2000/31/EC shall not apply* to the situations covered by this Article" (emphasis added).[118] In other words, the Directive created a new breed of obligations that are outside the scope of the safe harbour. These obligations are fairly similar to those that would be triggered by the non-application of the safe harbour to a hosting provider, as that provider would face liability for failing to have secured a license for copyrighted content it carries (and with the added twist that such license must also cover acts committed by is non-commercial users); however, this liability is subject to its own safe harbour if the provider has made best efforts to obtain the license, made best efforts to ensure the unavailability of copyrighted material for which the copyright owners have provided the relevant and necessary information, and acted

[116] Directive 2000/31/EC of the European Parliament and of the Council (2000), Article13.

[117] Sophie Stalla Bourdillon, 'Open Letter to the European Union' (*Medium*, 8 December 2016) https://medium.com/eu-copyright-reform/open-letter-to-the-european-commission-6560c7b5cac0.

[118] Directive (EU) 2019/790 of the European Parliament and of the Council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (2019) OJ L 130, Article 17 (3).

expeditiously upon notice to remove/disable access to a specific copyrighted work and made best efforts to prevent further uploads.[119]

The Article also designs a differentiated compliance regime for three types of entities: in addition to the online content-sharing providers who are subject to full compliance with the regime described above, *new* online content-sharing service providers the services of which have been available to the public in the Union for less than three years and which have an annual turnover below EUR 10 million will only have to comply with the duty of best efforts to obtain authorization and the notice and takedown obligation (no obligation with regard to future uploads), while those with an average number of monthly unique visitors of such service providers above 5 million will have to make best efforts to prevent further uploads (but not an obligation to do so following a notice).[120] It remains to be seen how this system will be implemented in practice, particularly given the duty of Member States to ensure the continued availability of non-infringing content[121] and to refrain from imposing general monitoring obligations.[122]

For now, the Directive has been challenged before the European Court of Justice by Poland, who requests the annulment of part of Article 17(4) claiming that the directive is shifting the responsibility of removing infringing uploads from the rightsholders onto platforms, who can only realistically do this by installing so-called "upload filters." The challenge is specifically grounded on the fundamental right to freedom of expression, arguing that an interference with such right is the unavoidable consequence of having a system of liability for failure to restrict content as opposed to no sanctions for unduly removing content.[123] The Court of Justice has yet to deliver its judgment, but Advocate General Øe has advised it to declare the compatibility of Article 17 (4)

---

[119] Ibid Article 17 (4).

[120] Ibid Article 17 (6).

[121] Ibid Article 17 (7).

[122] Ibid Article 17 (8).

[123] Paul Keller, 'CJEU Hearing in the Polish Challenge to Article 17: Not Even the Supporters of the Provision Agree on How It Should Work' (*Copyright Blog*, 11 November 2020) http://copyrightblog.kluweriplaw.com/2020/11/11/cjeu-hearing-in-the-polish-challenge-to-article-17-not-even-the-supporters-of-the-provision-agree-on-how-it-should-work/.

with the Charter of Fundamental Rights only insofar as the newly introduced duty of care applies to manifestly infringing content.[124] For other types of works, in turn, a judicial assessment will be required, given the sensitivity of these decisions and the platform's lack of independence in that regard.[125] He also made clear that providing safeguards for the availability of works due to copyright exceptions and limitations is an obligation of result, which naturally prevails over the obligations of effort imposed under Article 17(4).

The fight against terrorism is another issue area where the move towards enhanced responsibility of intermediaries has become apparent: first of all, with Directive 2017/541 that required Member States to take the necessary measures to ensure the prompt removal of online content constituting a public provocation to commit a terrorist offence. This provision is leading to two different types of implementations at the national level: new notice-and-takedown measures under the ECD, and criminal law measures allowing a prosecutor or a court to order companies to remove content or block content or a website, within a period of 24 or 48 hours.[126]

The recently adopted Terrorism Regulation[127] goes much further than the Directive, imposing on Hosting Services Providers (HSPs) a broader duty of care and proactive measures to remove terrorist content. For example, Article 3 provides that "competent authorities" will have the power to order a hosting service provider to remove "terrorist content" or disable access to it within one hour from the receipt of the order, while Article 6 prescribes that HSPs that are exposed to terrorist content must take specific measures to protect against its dissemination, such as appropriate technical and organizational measures to identify and remove it, easily accessible and user friendly mechanisms for users to report or flag it, and any other mechanisms to raise awareness over it and address its availability.

---

[124] Case C-401/09, *Poland v Parliament and Council* (2021) ECLI:EU:C:2021:613, para 198

[125] Ibid para. 218

[126] Directive (EU) 2017/541 of the European Parliament and of the Council on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA (2017) OJ L 88, 20-22.

[127] Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online.

Furthermore, HSPs will have to produce reports on such measures within 3 months of receiving a notification from a competent authority that the site is exposed to terrorist content, and may be required by the authority to "take additional necessary and proportionate measures." Seeking to reconcile these obligations with existing principles governing intermediary liability, the Regulation includes the caveats that the specific measures shall entail neither a general obligation to monitor content nor an obligation to seek facts or circumstances indicating illegal activity, and that there is no particular obligation to use automated tools (Article 5.8).

Similarly, to prevent possible friction with freedom of expression, the Regulation specifies that material disseminated to the public for educational, journalistic, artistic or research purposes or for the purposes of preventing or countering terrorism, including material which represents an expression of polemic or controversial views in the course of public debate, shall not be considered to be terrorist content (Article 1.3), and that the adopted measures shall be diligent, proportionate and non-discriminatory, taking into account the fundamental importance of the freedom of expression and information in an open and democratic society, with a view to avoiding the removal of material which is not terrorist content (Article 5.1). However, despite these caveats that were added to the original text of the proposal, concerns for freedom of expression remain given the very stringent timeline, the lack of judicial review and the absence of minimum standards for the appeal mechanisms that HSP are required to provide[128].

Finally, a third major area of responsibilization of internet intermediaries concerns the transmission of adequate audiovisual media content. In this respect, the reform of the Audiovisual Media Service Directive was proposed in 2016 and ultimately passed in 2018, with Directive 2018/1808. This legislation is perhaps the most paradigmatic shift of responsibilities to intermediaries that we have seen so far, as it extends to online platforms a large part of the obligations that traditionally applied to "linear" services. Specifically, the new Article 28a provides

[128] Joris von Hoboken, 'The Proposed EU Terrorism Content Regulation: Analysis and Recommendations With Respect to Freedom of Expression Implications' (2019) Working Paper from the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/EU_Terrorism_Regulation_TWG_van_Hoboken_May_2019.pdf.

that Member States "shall ensure that video-sharing platform providers take appropriate measures to protect:

- minors from programmes, user-generated videos and audiovisual commercial communications which may impair their physical, mental or moral development […];
- the general public from programmes, user-generated videos and audiovisual commercial communications containing incitement to violence or hatred directed against a group of persons or a member of a group based on any of the grounds referred to in Article 21 of the Charter of the Fundamental Rights of the European Union, or containing content the dissemination of which constitutes a criminal offence in the EU (namely child pornography or xenophobia);
- the general public from programmes, user-generated videos and audiovisual commercial communications containing content the dissemination of which constitutes an activity which is a criminal offence under Union law, namely public provocation to commit a terrorist offence […], offences concerning child pornography […] and offences concerning racism and xenophobia".

Relevant statutory provision	Protection from liability	Activity outside the scope of this protection	Likely implications for content moderation
CDA	**Providers or users of an interactive computer service (ICS)** cannot be treated as "publishers" of any information provided by another information content provider, and are not civilly liable for any action voluntarily taken in good faith to restrict or enable restriction of access to or availability of material that the provider or user	Intellectual property, privacy and criminal liability and sex trafficking law (§230 (e) (1)-(5)) Illegal content that the ICS contributes to create (FHC v Roommates) Illegal content that the ICS encourages (Jones v Dirty)	No active involvement in the creation of content No encouragement of illegal activity Warning of content or activities that may
E-commerce Directive	Hosting information society services (HISS) are not liable for the information stored at the request of a recipient of the service, on the following conditions: a) the provider does not have actual knowledge of illegal activity or information and, as regards claims for damages, is not aware of facts or circumstances from which the illegal activity or information is apparent; or b) the provider, upon obtaining such knowledge or awareness, acts expeditiously to remove or to disable access to the information.as long as they do not have actual knowledge of illegal activity or information, nor (as regard claims for damages) has awareness of facts and circumstances from which such illegality is apparent. c) the recipient of the service is not acting under the authority or the control of the provider. (art. 14 (1)- 14 (2))	Imposition by a court or administrative authority of obligation for the HISS to terminate or prevent an infringement (Art. 14 (3)) Imposition by Member States of duties of care on HISS which can reasonably be expected by them, and which are specified by national law, in order to detect and prevent certain types of illegal activities (Recital 48) Non-neutral role of HISS in the creation of content, in the sense that its conduct is not merely technical, automatic and passive, pointing to a lack of knowledge or control of the data which it stores (Google France) Promotion and optimization of the presentation by HISS of the offers for sale of a product (L´Oreal v Ebay) Awareness of facts and circumstances on the basis of which a "diligent economic operator" should have identified the unlawfulness of the user´s activity (L´Oreal v Ebay) Platform operator providing access to protected works with full knowledge of the infringing nature, by indexing torrent files which allow users to locate those works and to share them within the context of a peer-to-peer network (Stichting Brein v Ziggo) Video-sharing or a file-hosting &sharing platform contributing, beyond merely making that platform available, to giving access to such content to the public in breach of copyright: for instance, where it participates in selecting protected content illegally communicated to the public, provides tools on its platform specifically intended for the illegal sharing of such content or knowingly promotes (for instance, as demonstrated by the financial model) such sharing (Peterson v Google and Elsevier v Cyando)	Ability to detect and remove or disable access to content on the site No promotion or optimization of the presentation of content that might be illegal Minimal examination of facts and circumstances around third party content in the absence of notification No active role in content creation No design of the platform in a way that facilitates sharing and finding copyright-infringing content

Table 1: Content moderation implications of intermediary liability rules.

Relevant statutory provision	Responsability	Limitations and safeguards	Likely implications for content moderation
DSM Directive	**Online content-sharing providers (OCSSPs)** giving the public access to copyrighted material must: - obtain an authorization from right-holders which covers also the acts of users that do not act on a commercial basis or do not generate significant revenues (art. 17.1-17.2) - If no authorization is granted, make best efforts to obtain an authorization; and (b) make, in accordance with high industry standards of professional diligence, best efforts to ensure the unavailability of specific works and other subject matter for which the rightholders have provided the service providers with the relevant and necessary information; and in any event (c) act expeditiously, upon receiving a sufficiently substantiated notice from the rightholders, to disable access to, or to remove from their websites, the notified works or other subject matter, and make best efforts to prevent their future uploads in accordance with point (b). (art. 17. 4)	Application of art. 17 must not lead to general monitoring obligations (art. 17.8) Best efforts must be considered in light of the principle of proportionality, including the following elements: (a) the type, the audience and the size of the service and the type of works or other subject matter uploaded by the users of the service; and (b) the availability of suitable and effective means and their cost for service providers. (art. 17.5) The cooperation between OCSSPs and rightholders shall not result in the prevention of the availability of works or other subject matter uploaded by users, which do not infringe copyright and related rights, including where such works or other subject matter are covered by an exception or limitation (and in particular quotation, criticism, review and use for the purpose of caricature, parody or pastiche). (art. 17.7) OCSSPs must put in place an effective and expeditious complaint and redress mechanism that is available to users of their services in the event of disputes over the disabling of access to, or the removal of, works uploaded (art. 17.9)	Use of content recognition technologies that enable detection and prevention of uploaded content that matches the reference files provided by copyright owners Intensity of efforts primarily dependent on the size of the OCSSP and the target audience Creation of mechanisms that enable to take into account user´s view over applicable defenses and limitations prior to taking decision over legality of content Internal dispute resolution mechanism which does not guarantee independence, impartiality and statement of reasons for the decisions taken
AVMS Directive	**Video-sharing platform providers (VSPPs)** must take appropriate measures to protect: a) minors from programmes, user-generated videos and audiovisual commercial communications which may impair their physical, mental or moral development [...]; b) the general public from programmes, user-generated videos and audiovisual commercial communications containing incitement to violence or hatred directed against a group of persons or a member of a group based on any of the grounds referred to in Article 21 of the Charter of the Fundamental Rights of the European Union, or containing content the dissemination of which constitutes a criminal offence in the EU (namely child pornography or xenophobia) c) the general public from programmes, user-generated videos and audiovisual commercial communications containing content the dissemination of which constitutes an activity which is a criminal offence under Union law, namely public provocation to commit a terrorist offence [...], offences concerning child pornography [...] and offences concerning racism and xenophobia". (art. 28b (1))	Appropriate measures are "Without prejudice to Articles 12 to 15 of the E-Commerce Directive". (art. 28b (1)) Measures should be determined in light of the nature of the content in question, the harm it may cause, the characteristics of the category of persons to be protected as well as the rights and legitimate interests at stake, including those of the video-sharing platform providers and the users having created or uploaded the content as well as the general public interest. (art. 28 b (3)) Measures should be practicable and proportionate, taking into account the size of the video-sharing platform service and the nature of the service provided; they should not lead to any ex-ante control measures or upload-filtering which do not comply with article 15 of the E-Commerce Directive. (art. 28b (3)) Video-sharing platform providers must clearly inform users where programmers and user-generated videos contain audiovisual commercial communications, provided that such communications are declared as such ore the provider has knowledge of that fact. (art. 28b (2))	Adoption of various preventative measures, including those offered as examples in art. 28b (3): (a)including and applying in the terms and conditions of the video-sharing platform services the requirements referred to in paragraph 1; (b) including and applying in the terms and conditions of the video-sharing platform services the requirements set out in Article 9(1) for audiovisual commercial communications that are not marketed, sold or arranged by the video-sharing platform providers; (c) having a functionality for users who upload user-generated videos to declare whether such videos contain audiovisual commercial communications as far as they know or can be reasonably expected to know; (d)establishing and operating transparent and user-friendly mechanisms for users of a video-sharing platform to report or flag to the video-sharing platform provider concerned the content referred to in paragraph 1 provided on its platform; (e) establishing and operating systems through which video-sharing platform providers explain to users of video-sharing platforms what effect has been given to the reporting and flagging referred to in point (d); (f) establishing and operating age verification systems for users of video-sharing platforms with respect to content which may impair the physical, mental or moral development of minors; (g) establishing and operating easy-to-use systems (h) providing for parental control systems that are under the control of the end-user with respect to content which may impair the physical, mental or moral development of minors; (i) establishing and operating transparent, easy-to-use and effective procedures for the handling and resolution of users' complaints to the video-sharing platform provider in relation to the implementation of the measures referred to in points (d) to (h); (j) providing for effective media literacy measures and tools and raising users' awareness of those measures and tools.

T.Reg.	**Hosting Service providers** must disable access to terrorist content within one hour of receipt of an order by the competent authority. (art. 3.3) If exposed to terrorist content, they also must include in their terms and conditions and apply provisions to address the misuse of its services for the dissemination of such content to the public, and take specific measures to protect against this dissemination. (art. 5.1)	HSPs cannot be imposed obligations to monitor the information which they transmit or store, nor a general obligation actively to seek facts or circumstances indicating illegal activity (art. 5.8). Any requirement to take specific measures shall not include an obligation to use automated tools by the hosting service provider (art. 5.8). Specific measures must be: (a) effective in mitigating the level of exposure of the services of the hosting service provider to terrorist content; (b) targeted and proportionate, taking into account, in particular, the seriousness of the level of exposure of the services of the hosting service provider to terrorist content as well as the technical and operational capabilities, financial strength, the number of users of the services of the hosting service provider and the amount of content they provide; (c) applied in a manner that takes full account of the rights and legitimate interest of the users, in particular users' fundamental rights concerning freedom of expression and information, respect for private life and protection of personal data; (d) applied in a diligent and non-discriminatory manner (art. 5.3) Where specific measures involve the use of technical measures, appropriate and effective safeguards, in particular through human oversight and verification, shall be provided to ensure accuracy and to avoid the removal of material that is not terrorist content. (Art. 5.3) Material disseminated to the public for educational, journalistic, artistic or research purposes or for the purposes of preventing or countering terrorism shall not be considered to be terrorist content (Article 1.3) Adopted measures shall be diligent, proportionate and non-discriminatory, taking into account the fundamental importance of the freedom of expression and information in an open and democratic society, with a view to avoiding the removal of material which is not terrorist content (Art. 5.1)	Swift communication channels with competent authorities, including 24/7 availability of support staff Creation and staffing of terrorism task forces to identify and expeditiously remove terrorist content Enabling users to flag or report alleged terrorist content to the HSP Use of hash database of Global Counter Terrorism Forum, with information pooled voluntarily by various HSPs

Table 2: Content moderation implications of intermediary responsibility rules

The EU's proposal for a Digital Services Act, unveiled in December 2020, continues this trend of responsibilization especially with regard to what it calls "very large online platforms," i.e. online platforms which provide their services to a number of average monthly active recipients of the service in the Union equal to or higher than 45 million.[129] In particular, it requires such platforms to periodically identify, analyse and assess any significant systemic risks stemming from the functioning and use made of their services in the Union,[130] and put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified[131]. With regard to other online platforms and, more generally, providers of hosting services, the Act introduces a few procedural duties: the provision of tools enabling aggrieved

[129] Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/E (2020) COM/2020/825, Article 25.

[130] Ibid Article 26.

[131] Ibid Article 27.

parties to submit notices of illegal content; the duty to process these swiftly[132], and to prioritize notices received by "trusted flaggers;"[133] the duty to state the reasons underlying a removal or disabling of access for any particular content;[134] and the duty to provide access to an effective internal complaint-handling system,[135] as well as to engage in good faith with any certified out-of-court dispute settlement mechanism chosen by a consumer where the dispute could not be resolved through the internal complaint-handling system.[136] Finally, it imposes on all providers of intermediary services (including hosting, caching and conduit) the duty to include in their terms and conditions information on any restrictions that they impose (including policies, procedures, measures and tools used for the purpose of content moderation algorithmic decision-making and human review) in relation to the use of their service in respect of information provided by the recipients of the service;[137] and a general duty to act in in a diligent, objective and proportionate manner in applying and enforcing such restrictions, with due regard to the rights and legitimate interests (and fundamental rights) of all parties involved.[138]

In addition to the several proposals targeting the status of online platforms as intermediaries, antitrust law has been relied on to curb their power. Interestingly, antitrust intervention has been called for to fix societal problems that can relate to the exercise of freedom of expression, which is not traditionally considered in the hard core of antitrust law goals. The scathing report by the US House of Representatives in late 2020 noted that "news publishers raised concerns about the 'significant and growing asymmetry of power' between dominant online platforms and news publishers, as well as the effect of this dominance on the production and availability of trustworthy sources of news" and that "as a result, several dominant firms have an outsized influence over the distribution and monetization of trustworthy sources of news online, undermining the

[132] Ibid Article 14.

[133] Ibid Article 19.

[134] Ibid Article 15.

[135] Ibid Article 17.

[136] Ibid Article 18.

[137] Ibid Article 12 (1).

[138] Ibid Article 12 (2).

availability of high-quality sources of journalism."[139] In acknowledging these dangers, the Report went as far as to recommend breaking up online platforms so as to reduce their far-reaching power.[140]

Direct regulation has also been used in response to platforms' growing power. Australia is perhaps the most visible recent example, with its Digital Platform Inquiry, which forced Facebook to compensate publishers for their content shared on Facebook's platforms in an effort to make sure that pluralism, effective public discourse, and ultimately freedom of expression are safeguarded.[141] EU's proposed Digital Markets Act is yet another example of direct regulation that attempts to rein in online platforms by imposing a series of obligations to so called gatekeepers, which are defined as a provider of core platform services if (a) it has a significant impact on the internal market; (b) it operates a core platform service which serves as an important gateway for business users to reach end users; and (c) it enjoys an entrenched and durable position in its operations or it is foreseeable that it will enjoy such a position in the near future.[142] Facebook could well fall under this definition. While the DMA is not targeted at content moderation policies specifically, it forms part of the EU´s new digital package (2019-2024), whose overall aim is to ensure that digital markets remain open, transparent, and competitive.

5. The challenges of adequate content moderation

[139] US House of Representatives, 'Investigation of Competition in Digital Markets, Majority Staff Report and Recommendations' (2019), 62-63 https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf?utm_campaign=4493-519.

[140] Ibid 378.

[141] Mlex, 'Pushing Back at Big Tech' (*Mlex*, 15 March 2021) https://mlexmarketinsight.com//special-reports/pushing-back-on-big-tech-report.

[142] Proposal for a Regulation of the European Parliament and of the Council on Contestable and Fair Markets in the Digital Sector (Digital Markets Act), COM/2020/842 final, Article 3(1).

As Section 3 illustrates, we can observe a progressive mounting of pressure on digital intermediaries to fulfil broad responsibilities, in recognition of their key role in enabling online interactions and communications. The space provided by platforms like Facebook is increasingly compared to the one offered by public squares[143], private spaces with quasi-public function[144], or even to that of an essential facility[145]. This puts those platforms in the position of having a crucial impact on their users' enjoyment of fundamental rights.

The UNGPs,[146] a set of soft law standards for states and business enterprises adopted by the UN Human Rights Council in 2011, establish, under their second pillar, that business enterprises have a *responsibility to respect* human rights. The UNGPs make a distinction between the *state duty to protect* human rights, which reflect the international obligations that states have undertaken under international law, and the *business responsibility to respect,* which does not translate into an international obligation but indicates that "businesses should look to currently internationally recognised rights for an authoritative enumeration, not of human rights *laws* that apply to them, but of human *rights* they should respect."[147] This idea is rooted in the principle that business enterprises, when fulfilling their role of specialized organs of society performing specialized functions, are required to comply with all applicable laws and to respect human rights.[148]

The business responsibility to respect human rights means that businesses "should avoid infringing on the human rights of others and should address adverse human rights impacts with

---

[143] *Packingham v. North Carolina*, 137 S. Ct. 1730 (2017).

[144] See e.g. Daphne Keller, 'Who Do You Sue? - State and Platform Hybrid Power Over Online Speech" (2019) Hoover Institution Aegis Series Paper No. 1902 https://www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf.

[145] *Biden v Knight First Amendment Institute at Columbia University*, 593 US __ (2021).

[146] Guiding Principles on Business and Human Rights (n 10).

[147] Ruggie (n 11).

[148] Guiding Principles on Business and Human Rights (n 10) 1

which they are involved"[149] and is defined as a global standard of expected conduct[150] which exists notwithstanding where business enterprises operate, "independently of States' abilities and/or willingness to fulfil their own human rights obligations," and "over and above compliance with national laws and regulations protecting human rights."[151] The responsibility to respect also requires that business enterprises "avoid causing or contributing to adverse human rights impacts through their own activities, and address such impacts when they occur" and "seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts."[152]

However, this is an area that has remained blurred, due to the lack of articulation of binding human rights obligations for non-State actors. It is in light of this that our project endeavored to examine the concrete obligations that may be derived from existing law for the governance of platform content and activity. While, on the one hand, platforms' significance as a communication channel should weigh in favor of constraining their discretion to create their own rules, one cannot paint a picture that completely removes platforms' discretion in setting up their rules and standards as they see fit: their role as private regulators is preserved by right to property, the freedom to conduct business and, in certain cases, freedom of expression.

With regard to freedom of expression, platforms are generally at freedom to choose the type of content they host, as long as this does not involve violation of existing law. Although more than one US court has equated platforms like Facebook to a public square,[153] no case or law has so

---

[149] Guiding Principles on Business and Human Rights (n 10) Guiding Principle 11.

[150] Ruggie (n 11) 13-14.

[151] Guiding Principles on Business and Human Rights (n 10) 13.

[152] Guiding Principles on Business and Human Rights (n 10) Guiding Principle 13.

[153] See *Packingham* (n 143) (prohibiting the government from banning sex offenders from it entirely as that would violate the well-established general rule that the Government may not suppress lawful speech as the means to suppress unlawful speech); *Marsh v. Alabama* 326 U.S. 501 (1946) (upholding a speaker's right, under the First Amendment, to distribute religious literature within the defendant's "company-owned town", which is where "a private entity owns all the property and controls all the municipal functions of an entire town"); *PruneYard Shopping Center v. Robins* 447 U.S. 74 (1980) (upholding the right of speakers to gather signatures for a petition in a privately owned shopping center).

far held that they are required (like a public square) to accommodate all lawful speech. This is especially the case where alternative forums exist,[154] as, in principle, that would mean that users can choose between those alternatives depending on the alignment with their core values and norms. That libertarian assumption underpins the argument advanced, in the very early days of the Internet, by Johnson and Post: they suggested that cyberspace would involve a shift away from State (and territorial) sovereignty, where users would primarily obey the laws of different electronic entities.[155]

Since then, however, we have learned that this libertarian assumption has limits: first of all, we cannot simply accept as a dogma the fact that people vote with their feet based on their understanding of the community norms. As several studies have pointed out, users do not necessarily read or understand the terms of service.[156] Further, even if they dislike a particular rule or change of the community standards, they may be unlikely to switch in the presence of network effects and switching costs: they have connections in those communities, and they may have developed content or habits that cannot be simply exported to the new environment. Finally, the moderation practices developed by users may have strong speech rights or associational interest that deserve protection, even where in conflict with the platform's norms.[157]

But see also *Johnson v. Twitter Inc*., No. 18CECG00078 (California Superior Court, 2018) (California Superior Court refusing to consider Twitter akin to a 'private shopping mall' that was 'obligated to tolerate protesters'); and *Prager University v. Google LLC*, No. 18-15712 (N.C.D.C., 2018) (Northern California District Court refusing to see YouTube as a state actor in accordance with the 'public function' test, arguing that providing a video sharing platform fulfils neither an exclusive nor a traditional function of the state), affirmed on appeal in *Prager University v. Google* LLC, 2020 WL 913661 (9th Cir., 2020).

[154] The case-law of the ECHR found as much in a case involving a shopping center in the United Kingdom: see *Appleby and Others v. the United Kingdom,* no. 44306/98, ECHR 2003-VI.

[155] David R. Johnson and David Post, 'Law and Borders: The Rise of Law in Cyberspace' (1996) 48 Stanford Law Review 1367.

[156] David Berreby, 'Click to Agree With What? No One Reads Terms of Service, Studies Confirm' *The Guardian* (3 March 2017).

[157] Nicolas Suzor, 'The Rule of Law in Virtual Communities' (2010) 25 Berkeley Technology Law Journal 1817.

At the same time, it cannot be sustained that moderation is an optional feature for platforms, given the need to prevent the dissemination of low-quality information, manipulated information and abuse.[158] This presents public regulators with a dilemma: in order to achieve public policy goals, they need to determine what is the appropriate degree of oversight on the platform's rules and practices. A dilemma is also faced by platform content moderators about how to structure their internal decision-making, given the need to exercise quick judgments on vast amounts of content that is produced instantaneously, and may raise several complex legal and socio-economic issues. Typically, this has been addressed by relying on 3 different approaches: (1) Artisanal approaches, which rely on teams from 5 to 200 staff members; (2) Community-reliant approaches, which typically combine formal policy made at the company level with volunteer moderators; and (3) Industrial approaches, where thousands of workers are employed to enforce rules made by a separate policy team.[159] In this third type of approach, that is followed by very big platforms like Facebook and YouTube, the risk of losing relevant context is greater, due to the greater use of artificial intelligence to make decisions.[160]

The problems of "overblocking" generated by artificial intelligence have been noted by several commentators, lamenting the excessive reliance by both companies and regulators on the promises offered of scale and efficiency of algorithmic moderation processes, thereby discounting the downsides and shortcomings of these processes. The criticism is not about the use of algorithmic moderation *per se*, but about the lack of safeguards designed to prevent adverse impact on freedom of expression: for instance, current algorithmic moderation practices can be improved through a recognition of the importance of specific contextual elements, such as the identity of the speaker and that of the receiver of a message.[161] Similarly, the inclusion of a wide variety of sources and languages in the datasets used to train algorithmic systems would be a

---

[158] James Grimmelmann, 'The Virtues of Moderation' (2015) 17 Yale Journal of Law & Technology 42.

[159] Robyn Caplan, 'Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches' (*Data & Society*, 14 November 2018).

[160] Ibid.

[161] Emma Llansó, Joris van Hoboken, Paddy Leerssen and Jaron Harambam, 'Artificial Intelligence, Content Moderation, and Freedom of Expression' (*Transatlantic Working Grou*p, 26 February 2020) https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf.

significant step forward in taking the interests of diverse communities into account when designing algorithmic moderation processes. Furthermore, "human in the loop" mechanisms can preserve accountability and contestability of automated decisions that may otherwise work undeterred, perpetuating errors of both under and over-inclusion.

To understand how this may occur in practice, it is useful to distinguish between two different types of technologies used for content moderation: matching and classification. In the former, new content is compared and contrasted with files in a database of prohibited content, to possibly filter out any match (which can be even in percentile terms, such as 80%); whereas in the latter, artificial intelligence (typically, supervised machine learning) is used to classify or predict content as belonging to one of several categories.[162] Some of the strongest criticisms, like the one about the possible bias of learning datasets, are particularly relevant in this second context, which is crucial because it is where the "cooking recipe" for content moderation tools is made: the more limited and unrepresentative the sample, the more moderation actions will address the concerns of only a subset of the population. However, representativity in the sense of comprehensiveness of the reference database is an equally valid concern with regard to hashing, as it helps avert disproportionately adverse effects for less well-known or disadvantaged communities. Secrecy in the procedures followed to establish these databases and how the information is exchanged between different players are particularly acute for extremist content, which is now primarily moderated through action taken by members of the Global Internet Forum to Counter Terrorism. In 2016, Twitter, Facebook, Google and Microsoft announced a shared database of hashes for this type of content, developed without public scrutiny at any stage, which explain why commentators have been referring to these instances of cooperation as "content cartels."[163]

---

[162] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2019) 7 Big Data & Society (online) https://journals.sagepub.com/doi/10.1177/2053951719897945.

[163] Evelyn Douek, 'The Rise of Content Cartels' (*Knight First Amendment Institute*, 2020) https://knightcolumbia.org/content/the-rise-of-content-cartels.

6. Methodology

The project was divided in three stages: (a) data collection, (b) data curation, (c) data analysis. By data, we mean the body of documents that comprise Facebook's content policies and the body of documents produced by international organisations that concern freedom of expression. The goal of the project was to juxtapose the guidance provided by the international community with Facebook's content policies all in the context of the transformation of freedom of speech online and the role of social media platforms and with the view to determine Facebook's handling of freedom of expression given the available guidance. We do not assess Facebook's enforcement of its content policies.

a) Data collection

i. Facebook

Facebook's content policies have evolved over time and today comprise multiple documents. In our analysis we included the Terms of Service, the Community Standards, the Advertising Policy, the Code of Conduct, the Privacy Policy, the Pages Terms, and the Facebook Live Policies. The majority of the provisions that were relevant for our purposes were found in the Community Standards and the Terms of Service.

To access historical versions of the above documents we relied on the Wayback Machine (https://web.archive.org), which has saved copies of Facebook's policies from late 2005 onwards. We aimed to review every policy change Facebook made in all of the listed documents. We set the interval between versions at two weeks, reasoning that no meaningful change could be introduced and then amended in a matter of only one month (two plus two weeks). If a version was unavailable we moved to the next available one even if it was not two weeks later. Among the reasons why versions were unavailable are broken links, and lack of stored versions in

English. For missing versions, we attempted to consult Facebook's own archive[164], but that was also often inaccessible despite our efforts to access it through different browsers and different locations (UK, Switzerland, Italy, and Brazil). Our review period spanned November 2005 to November 2020.

ii. International community

We aimed to collect all international instruments that concern at least partially freedom of expression. This included binding and non-binding instruments; instruments that are specifically concerned with freedom of expression; instruments where freedom of expression is just one of the rights addressed; and instruments that are concerned with other rights, but which include provisions that have a significant bearing on how freedom of expression is exercised (e.g. Convention on the Prevention and Punishment of the Crime of Genocide). We recognize that binding instruments are binding only upon the member states of the issuing international organization, and that most non-binding instruments are still addressed to member states and not to private corporations such as Facebook. Nevertheless, as explained earlier,[165] in light of the fact that business enterprises still hold a responsibility to respect human rights under the UNGPs, we see the rights, obligations, and accompanying interpretations enshrined in these documents as best practices, guidance, and standards that even private corporations can and should aspire to. That said, we only included instruments where the relevant right or obligation, if exercised or enacted by states or corporations, would have private individuals as their subject. This would normally exclude obligations to criminalize conduct (which are plentiful in international instruments), since these are addressed to states. However, Facebook states in its ToS that it can remove content that is "unlawful" and in its Transparency Center it states that "when something on Facebook or Instagram is reported to us as violating local law, but doesn't go against our Community Standards, we may restrict the content's availability in the country where

[164] See https://www.facebook.com/communitystandards/recentupdates/all_updates/.

[165] See Sections 1 and 5.

it is alleged to be illegal."[166] Moreover, the Bylaws of Facebook's Oversight Board state that the Board does not review cases "where the underlying content is criminally unlawful in a jurisdiction with a connection to the content."[167] We therefore concluded that international provisions that ask for conduct criminalization can inform and affect Facebook's policies even if, technically, Facebook could never be the recipient of such obligations.

Our covered period spans 1948 to 2020, we catalogued 48 international instruments, and a total of just short of 400 provisions across them ('All International Instruments' tab in the dataset).[168]

b) Data curation

i. Facebook

Because the project required us to track the revisions Facebook made to its content policies, we catalogued both new/original provisions that we concluded concerned freedom of expression, and the ensuing revisions, excluding formatting and stylistic changes, as well as all changes that we concluded did not concern freedom of expression. To identify changes from one version to another, we used Microsoft Word's automatic comparison tool. We took an expansive interpretation of areas that concern freedom of expression, which we grouped as follows:

- Chilling effect factors: Anonymity, Relations with governments;
- Types of expression and areas of freedom of speech: Hate speech (including antisemitism), terrorism, bullying and harassment, nudity, protected characteristics and discrimination;

[166] Facebook Content Restrictions Based on Local Law https://transparency.fb.com/data/content-restrictions/.

[167] Oversight Board Bylaws, Article 2, Section 1.2.2.

[168] Dataset (n 9).

- Remedies and redress mechanisms;
- Special rules for fake news and misinformation;
- Special rules for the protection of children;
- Intellectual property limitations and access to knowledge;
- Stakeholder involvement in shaping freedom of expression policies.

Within those areas we catalogued new/original provisions and their revisions, which affect the freedom of expression rights or obligations of Facebook toward users, or of users toward Facebook, or of users toward each other. In total, we catalogued 223 original provisions and revisions across Facebook's content moderation policies.

ii. International community

The project required us to compare Facebook's policies on freedom of expression and the available guidance from the international community. As Facebook was opened to the public in 2004, we treated the guidance provided by the international community before and after Facebook's founding separately. Rules and guidance that predated Facebook's founding could have been taken into account even in the first version of Facebook's content policies. For rules and guidance that came out after 2005, we compiled a list of 34 milestone provisions, which we used to focus our analysis ("Post-2005 Instr. Milestones" tab in the dataset).[169] The milestone provisions are not the only ones we considered in our analysis; they were just a helpful focusing device.

c. Data analysis

For the data analysis, we consulted the relevant literature and we chronologically juxtaposed Facebook's changes to its content policies with the evolving rules and guidance that was

[169] Ibid.

becoming available over time at the international level. The literature review helped us develop our initial focus on the most contentious issues, but we expanded from there into new areas that may have received less attention in news coverage and the scholarly debate (e.g. IP limitations, remedies, Facebook's relationship with governments etc.). We tracked milestone changes in Facebook's policies and we linked them to relevant rules and guidance by the international community to monitor compliance (see "Notes" column in the annex).[170] We then extracted high-level patterns and insights, which we analyze in this report in Part II.

[170] Ibid.

PART 2: FINDINGS, OBSERVATIONS, AND RECOMMENDATIONS

While social media companies have received a great deal of criticism for their policies, one should recognize that policies are not developed in a vacuum. Rather, corporate policies are molded within the boundaries of regulation and under the guidance that is available on the areas that policies touch on. It would be non sequitur to require private companies to respect and uphold human rights to a high standard, when those that are primarily entrusted with developing such standards remain silent or confused on the matter. And vice versa, when rules and guidance are available, compliance becomes more imperative, and deviation becomes a matter of deliberate choice, rather than justifiable ignorance.

We analyze below Facebook's content moderation policies under the light of the standards, guidance, and recommendations developed at the international level. As explained in Part 1, while Facebook is not technically bound by international law, compliance with the authoritative (and jurisdiction-agnostic) mandates and guidance drafted by the international community demonstrates respect for the rule of law and for the well-being of Facebook's over 2 billion users, who operate under the private ordering regime crafted and singe-handedly managed by Facebook.

1. Facebook's content moderation policies developed slowly, but in part so did the guidance by the international community

Overall, the picture that emerges is one of slow and insufficient response to the challenges of content moderation and freedom of expression on online platforms, without that meaning that there are no bright spots.

By the time Facebook was founded, the international community had had ample time to develop sufficient standards on freedom of expression, which Facebook could have taken into account early on. While this was not always the case, we did find that in many areas guidance was indeed sufficient. We also found that the international community exhibited good reflexes in

some emerging issues online such as fake news. However, on other contemporary issues, like the role of anonymity online, the role of social media in spreading terrorist content, or the compliance of social media platforms with government requests for user data, the international community reacted slowly or inadequately despite the lead time it had before those issues became of critical concern.

Overall, Facebook did not seem to prioritize detailed content moderation in its early days, as it did not make best use of tools and guidance available even at the time of its founding. Facebook's content moderation policies improved vastly in the period 2018-2020, but various weaknesses, like, for instance, categorical prohibitions of nudity, or the name policy that amounted to an effective ban on anonymity could have been avoided from the outset under standard freedom of expression doctrines that pre-dated Facebook. Generally, though today Facebook has sufficiently good content policies in place in a few areas, such as bullying and fake news,[171] Facebook often did not catch up in time with international guidance where available, and in many other areas, where the international community had left a gap, it missed an opportunity to spearhead the drafting of good policies. We develop these insights in detail in Section 6.

[171] Note that in this study we do not assess the enforcement of Facebook's policies, only the policies themselves. We appreciate that Facebook's handling of fake news has attracted negative attention, but those criticisms are usually aimed at how Facebook implements its policies, not the policies themselves. See below Section 6.A.

Area of Freedom of Expression	International Community Response	Facebook Response
Terrorism	Vague	Late/vague
Remedies	Early	Late/Inadequate
Anonymity	Late	Restrictive
False/Fake news	Timely	Timely
Hate speech	Late	Late/vague
Protection of children	Early	Late
Protected characteristics	Adequate	Late
Nudity	N/A	Restrictive
Bullying	N/A	Adequate
Transparency on Government Requests	Late	Inadequate
IP and access to knowledge	Generic	Restrictive

Table 3: Summary of findings on Facebook's response to areas and determinants of freedom of expression compared to the international community

2. The disconnect between social media platforms and the international community

Perhaps the biggest obstacle in international freedom of expression standards to influence social media platforms is the disconnect between them. By and large, international law is a states' game: the laws, standards, and guidance are developed by states for states. This means that,

technically speaking, social media platforms, such as Facebook, are not bound by the work done under the auspices of international organizations. Instead, once states transplant their international commitments into national law, only then companies are required to follow what national law mandates (with the exception of EU law that can directly bind natural and legal persons too).

Traditional international human rights law considers states as the primary bearers of human rights obligations. The notion of international human rights law having a 'special character' refers to the idea that, contrarily to normal international law obligations, human rights law obligations are concerned with the duties that states have towards *individuals:* while a traditional treaty usually creates rights and obligations *vis-à-vis* other states (usually excluding other actors), human rights treaties create rights whose beneficiaries are individuals.[172] In line with this conception, most of the interpretive guidance offered by human rights monitoring and implementation mechanisms is thus addressed to states: for example, both UN Treaty Bodies and Special Procedures are meant to monitor states' compliance with their international obligations and to offer them guidance on their implementation.

The UN Guiding Principles differentiate between the States' duty to protect and the corporate responsibility to respect. The latter is based on a near-universal recognition that corporations have a responsibility to respect human rights.

International law, however, is not completely foreign to the issue of business enterprises and their impact on the enjoyment and protection of human rights. As mentioned, in 2011, the Human Rights Council adopted the UNGPs,[173] a set of soft law standards for states and corporations.

[172] Fréderic Mégret, 'Nature of Obligations' in Daniel Moeckli, Sangeeta Shah, and Sandesh Sivakumaran (eds), *International Human Rights Law* (3rd ed, Oxford University Press, 2017) 88–89.

[173] Human Rights Council, 'Guiding Principles on Business and Human Rights - Implementing the United Nations' "Protect, Respect and Remedy" Framework,' HR/PUB/11/04 (2011).

The UNGPs do not create new international law obligations, nor do they limit or undermine the obligations that states have undertaken under international law. Rather, they are to be understood "as a coherent whole and should be read, individually and collectively, in terms of their objective of enhancing standards and practices with regard to business and human rights so as to achieve tangible results for affected individuals and communities, and thereby also contributing to a socially sustainable globalization." [174] They rest on three pillars: (1) states have a duty to protect against human rights abuses by third parties, including businesses, by enacting appropriate policies, regulation and adjudication; (2) corporations have the responsibility to respect human rights, including acting with due diligence to avoid infringing the rights of others and to address adverse human rights impacts; (3) access to effective remedy, both judicial and non-judicial, should be granted to victims.[175] The UNGPs differentiate between the states' *duty to protect* (which is derived from international law obligations), and the corporate *responsibility to respect*. The latter is based on a near-universal recognition that corporations have a responsibility to respect human rights.[176]

Post-1990, when the Internet became commercialized, fewer than ten initiatives on freedom of expression by international organizations are either directly addressed to or involve online intermediaries, such as ISPs or social media platforms, and of those, half simply recognize the role of online intermediaries without however providing substantial guidance.

[174] Ibid.

[175] John Ruggie, 'Global Governance and "New Governance Theory": Lessons from Business and Human Rights' in Alynna J. Lyon, Kendall Stiles, Alistair Edgar, Kurt Mills, and Peter Romaniuk (eds), *Global Governance: A Review of Multilateralism and International Organizations* (Brill Nijhoff, 2014) 7.

[176] John Ruggie, 'The Social Construction of the UN Guiding Principles on Business & Human Rights' (2017) HKS Faculty Research Working Paper Series RWP17-030 http://www.hks.harvard.edu/publications/social-construction-un-guiding-principles-business-human-rights, 13-14.

Therefore, there is a concomitant expectation that international instruments and organizations provide guidance to online platforms and other digital intermediaries in hopes that they will voluntarily comply. Considering the global reach and operation of social media companies, the disconnect between work done at the international level and social media companies operating within the confines of national laws becomes particularly problematic. While there are national differences in the protection and promotion of freedom of expression, there exists a core of protections and expectations that seems universal. At the very least, then, one would expect the international community to engage global social media platforms regarding a minimum level of protections.

Our research indicates that, post-1990, when the Internet became commercialized, fewer than ten initiatives on freedom of expression by international organizations are either directly addressed to or involve online intermediaries, such as ISPs or social media platforms, and of those, half simply recognize the role of online intermediaries without however providing substantial guidance.[177] The bulk of lawmaking, quasi-lawmaking and accompanying guidance at the international level which we catalogued at just over 25 instruments for that period, engages only states and it reaches online platforms through trickling down from international organizations to national governments. Activity directly engaging online intermediaries starts around 2009, approximately a year after the official adoption of the Ruggie Framework on Business and Human Rights.[178] However, it is slow to pick up the pace, with most work having been done only in the past five years (post-2016).

This recently intensified engagement of the international community with online intermediaries is certainly welcome, but with the pervasive role of intermediaries known since the mid-90s already, the long delay in directly addressing and guiding them, as well as in raising expectations around their operation is somewhat disappointing. While one can surely complain about the poor practices of social media platforms, it is worth considering whether available guidance has been available to them at the international level at which they operate, even if they wanted to adopt best practices.

---

[177] Dataset available at https://doi.org/10.5518/1072.

[178] John Ruggie, 'Protect, Respect and Remedy: A Framework for Business and Human Rights' (2008) A/HRC/8/5.

In 2009 the Safer Social Network (SSN) principles is the first major initiative driven by states but also engaging online platforms.[179] SSN was the result of discussions in the Social Networking Task Force set up by the European Commission in April 2008, which involved social networking sites, NGOs and researchers and it represented the first attempt for comprehensive regulation sponsored or heavily supported by online platforms, including Facebook. As a first step in the direction of involving tech giants, it is no surprise that it was a non-binding self-regulatory initiative, which nonetheless was labeled as "a good example of industry self-regulation, an approach favored by the Commission if effectively implemented."[180] The initiative was part of a wider discussion led by the European Commission, which was reviewing protection of minors online from such risks as grooming and cyber-bullying as part of the objective set by the Digital Agenda for Europe to enhance trust on the Internet.

In 2013, the Report of the Inter-American Commission on Human Rights Special Rapporteur on Freedom of Expression and the Internet came out as the second major instrument to provide non-binding guidance to not just governments but also civil society and non-state actors (including social network platforms) "in order to clear the way for this conceptually and technically new territory, and stimulate the revision and adoption of legislation and practices in view to achieving the full realization of the right to freedom of thought and expression through the Internet."[181] Among others, the report shed light on the importance of anonymity as a safeguard for a proper exercise of freedom of expression, and recommended that content moderation be rights-compatible, that social network sites Terms of Service should be transparent, clear, accessible and consistent with international human rights law and that SNSs should be transparent in disclosing governmental requests for content takedowns. Many of these provisions are at the heart of Facebook's controversies over the years.

The sudden propelling of fake news in the spotlight during the 2016 US Presidential Elections and the prominent role social media platforms, and Facebook in particular, played, finally left no

---

[179] Press Statement, 'Social Networking: Commission Brokers Agreement Among Major Web Companies,' IP/09/232 (*European Commission*, 10 February 2009).

[180] Ibid.

[181] Catalina Botero Marino, 'Report of the Inter-American Commission on Human Rights Special Rapporteur on Freedom of Expression and the Internet' (2013) OEA/Ser.L/V/II, para 3.

margin for the international community to involve private intermediaries. The 2017 Joint Declaration of Special Rapporteurs on Fake News[182] included detailed guidelines addressed to states, but also to intermediaries, journalists and media outlets.[183] These concerned both the clarity of content policies, but also redress mechanisms and due process. While the Joint Declaration is not binding, the comprehensive and immediate coverage it provided on a hot topic showed good reflexes on the side of the international community.

By 2018 it becomes apparent that internet intermediaries of all kinds hold pervasive power across all aspects of freedom of expression. Unlike some of the previous instruments, the Council of Europe Recommendation of the Committee of Ministers on the Roles and Responsibilities of Internet Intermediaries provides comprehensive and general guidance on various aspects of safeguarding freedom of expression stating that "Internet intermediaries should in all their actions respect the internationally recognised human rights and fundamental freedoms of their users and of other parties who are affected by their activities."[184]

The CoE Recommendation is followed by the 2018 Report of the UN Special Rapporteur on the Promotion and Protection of the Rights to Freedom of Opinion and Expression, which recognises that "Internet companies have become central platforms for discussion and debate, information access, commerce and human development. … Few companies apply human rights principles in their operations, and most that do see them as limited to how they respond to government threats and demands. However, the UNGPs establish 'global standard[s] of expected conduct' that should apply throughout company operations and wherever they operate. While the UNGPs are non-binding, the companies' overwhelming role in public life globally argues strongly for their adoption and implementation."[185] The 2019 Report on the Promotion and Protection of the Right to Freedom of Opinion and Expression doubles down on some of those ideas, but does not provide any substantial new guidance, and the 2019 Report of the Special Rapporteur on

[182] UN, OSCE, OAS, ACHPR Special Rapporteurs 'Joint Declaration of Freedom of Expression and "Fake News", Disinformation and Propaganda' (March 2017).

[183] Ibid, Sections 4 and 5.

[184] Ibid para 2.1.1.

[185] Ibid paras 9-10.

freedom of religion specifically calls social media platforms to "enforce terms of service and community rules that do not allow the dissemination of hate messages, provide more transparency in their efforts to combat cyberhate and offer user-friendly mechanisms and procedures for reporting and addressing hateful content."[186]

A few other documents marginally involve online platforms, but not in a substantial way, and so they cannot be counted as providing material guidance. The 2017 Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression expands the interest of the international community to private online intermediaries, but excludes platforms like Facebook—instead it focuses on connectivity providers. While this is a welcome recognition of the private power held by various types of intermediaries, not just of social media platforms, it stopped short of considering the complexities of platforms such as Facebook. Along similar lines, the 2013 General Recommendation 35 on Combating Racist Hate Speech, issued by the UN Committee on the Elimination of Racial Discrimination (CERD), in clarifying the definition of racist hate speech, explicitly mentions that social media should adopt guidelines incorporating CERD principles and other fundamental human rights, as hate speech policies have been deemed to lack clarity and to be enforced inconsistently.[187] General Recommendation 35 was not a real engagement of social media platforms, but rather a recognition of their power and the urgency of having them respect international human rights standards. A few other international instruments also generally developed the concept of corporate social responsibility, but, again, these were general calls for corporations to act responsibly, not constitutive documents of expectations of or even obligations toward their users.[188]

---

[186] Ibid para 88.

[187] Ibid para 39.

[188] Maud de Boer-Buquicchio, 'Report of the UN Special Rapporteur on the Sale of Children, Child Prostitution and Child Pornography' (2014) A/69/262; Rita Izsák, 'Report of the UN Special Rapporteur on Minority Issues" (2015) A/HRC/28/64; David Kaye, 'Report of the UN Special Rapporteur on the Promotion and Protection of the Rights to Freedom of Opinion and Expression' (2016) A/71/373; David Kaye, 'Report of the UN Special Rapporteur on the Promotion and Protection of the Rights to Freedom of Opinion and Expression' (2017) A/72/350; David Kaye, 'Report of the UN Special Rapporteur on the Promotion and Protection of the Rights to Freedom of Opinion and Expression' (2019) A/HRC/41/35; Maud de Boer-Buquicchio, 'Report of the UN Special Rapporteur on the Sale and Sexual Exploitation of

3. The 'take-it-or-leave-it' nature of content policies and the overlooked 'legality, necessity, proportionality' standard

The take-it-or-leave-it nature of platforms' online terms of service has always been problematic. Facebook, as early as the first version of its Terms of Service, told users "You understand and agree that Facebook may review and delete or remove any Member Content that in the sole judgment of Facebook violate this Agreement or which might be offensive, illegal, or that might violate the rights, harm, or threaten the safety of Members."[189] Evidently, the exercise of users' freedom of expression right online has always been dependent on the judgement and good will of online intermediaries. This kind of digital 'constitutionalism', to use the words of Suzor,[190] and the concomitant power over the rights of netizens, justifiably raises concerns but also expectations that the power of the platforms will be exercised with reasonableness, fairness, and predictability, reminiscent of those attached to state actors.

Virtually all human rights, including the right to freedom of expression, are not absolute, but subject to limitations acknowledged in the human rights treaties themselves. However, in turn, these limitations are not arbitrary, but rather follow a 'legality, necessity, proportionality' standard. In that sense, the international community has always provided the means to online platforms to steer away from black and white rules and rather take a more nuanced approach to users' freedom of expression on their platforms.

Children, Including Child Prostitution, Child Pornography and Other Child Sexual Abuse Material' (2020) A/HRC/43/40.

[189] Dataset available at https://doi.org/10.5518/1072.

[190] Nicholas Suzor, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4 Social Media + Society (online).

One of the hallmark contributions of international human rights law is precisely the development of a framework that carefully balances the protection of rights with the narrowly tailored restrictions that could be imposed. Virtually all human rights, including the right to freedom of expression, are not absolute, but subject to limitations acknowledged in the human rights treaties themselves. However, in turn, these limitations are not arbitrary, but rather follow a "legality, necessity, proportionality" standard, which has been developed long before Facebook was created, but was fleshed out in more detail for the online platform context later on. In that sense, the international community has always provided the means to online platforms and social media networks to steer away from black and white rules and rather take a more nuanced approach to users' freedom of expression on their platforms. It is, of course, more resource-intensive for platforms to develop and enforce nuanced rules, but one should not confuse the intentional downgrading of adequate protection of users' rights in favor of economization of resources with lack of proper guidance if one wanted it. And while it is also true that black-and-white rules are clearer than "proportionality" rules, it is only the latter than can result in speech maximization, if that is the priority.

Art. 19 of the International Covenant on Civil and Political Rights (ICCPR) recognizes that freedom of expression can be legitimately restricted for the protection of the rights and reputation of others, of national security or public order, or of public health or morals. Additionally, art. 20 of the ICCPR prohibits any propaganda for war and any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. However, it is not sufficient that a restriction pursues one of these legitimate aims: for a limitation to be compliant with human rights law, it must also be (1) prescribed by law (legality); (2) necessary to pursue the legitimate aim(s) identified by the provision; (3) proportionate to the legitimate aim(s).

The tripartite test does not offer a ready-made recipe: each case needs to be carefully assessed and there could be different answers that would still be compliant with the test. Nonetheless, the application of the tripartite test to the right of freedom of expression has been clarified in many instances.

Setting aside the extensive guidance provided by case-law, already in 1995 the then Special Rapporteur on the promotion and protection of Freedom of Expression had clarified in his Report

the nature and scope of the right to freedom of opinion and expression and the restrictions and limitations to the right to freedom of expression.[191]

Subsequent reports clarify the scope of the right and its application with respect to different thematic areas. In 2011, the Special Rapporteur dedicated an entire report to the right to freedom of opinion and expression exercised through the Internet.[192] The Human Rights Committee's General Comment 34 also provides extensive guidance on the scope and application of Article 19: it addresses in detail the application of the tripartite test (legality, necessity, proportionality) and offers guidelines for interpreting Article 19 of the ICCPR in light of current contexts, clarifying in particular the legality of restrictions, including blasphemy laws, "memory" laws, treason, counter-terrorism, lèse-majesté, defamation of the head of state and the protection of honor of public officials.[193] When addressing the issues of electronic information dissemination systems, the General Comment clarifies that any restrictions to their operations must be compatible with Article 19(3). Lastly, in his 2018 Report, the Special Rapporteur on the promotion and protection of freedom of expression called social media platforms to take a "human rights by default" approach to content moderation, stating that:

> ***"Terms of service should move away from a discretionary approach rooted in generic and self-serving 'community' needs. Companies should instead adopt high-level policy commitments ... in a manner consistent with human rights law. … Companies should incorporate directly into their terms of service and 'community standards' relevant principles of human rights law that ensure***

[191] Abid Hussain, 'The Nature and Scope of the Right to Freedom of Opinion and Expression, and Restrictions and Limitations to the Right to Freedom of Expression' (1995) E/N.4/1995/32.

[192] Frank La Rue, 'Promotion and Protection of the Right to Freedom of Opinion and Expression' (2011) A/66/290.

[193] General Comment No. 34 on the Right to Freedom of Opinion and Expression (2011) CCPR/C/GC/34.

content-related actions will be guided by the same standards of legality, necessity and legitimacy that bind State regulation of expression."[194]

The Report provides specific guidance to social media companies on how to apply the tripartite test to their activities. Even more detailed guidance on the application of the test by social media platforms is provided in the 2019 Report of Special Rapporteur on the promotion and protection of freedom of expression which focuses on hate speech.

The Report of the UN Special Rapporteur on the Promotion and Protection of the Rights to Freedom of Opinion and Expression (2011) specifically addressed the use of the tripartite proportionality test in an online context, and unsurprisingly, the proportionality test (or lack thereof) has been consistently challenged in the decisions of Facebook's Oversight Board.

Facebook has dramatically improved its content policies in many areas of freedom of speech, such that the proportionality standard is already embedded in the rules (it is a different question whether the balance has been struck correctly). What is important to note here is not that adding nuance to policies sometimes comes late, but rather, that until platforms develop nuanced policies underpinned by the necessary due process, black-and-white policies do not be the alternative default. As mentioned previously, the triptych of "legality, necessity, proportionality" used to assess limitations on freedom of speech has been a hallmark of international lawmaking for several decades now. Facebook does not introduce qualifications into content moderation until the creation of Community Standards in 2010 and it is not until 2012 that we begin to see language reminiscent of a proportionality standard (Community Standards version of December 2012: "We understand that graphic imagery is a regular component of current events, but must *balance* the needs of a diverse community") (emphasis added). The Report of the UN Special Rapporteur on

[194] David Kaye, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (2018) A/73/348.

the Promotion and Protection of the Rights to Freedom of Opinion and Expression (2011) specifically addressed the use of the tripartite proportionality test in an online context, and unsurprisingly, the proportionality test (or lack thereof) has been consistently challenged in the decisions of Facebook's Oversight Board.

Due to its role as a global virtual "public sphere,"[195] which has become essential in not just social connections, but democratic discourse, Facebook has an elevated duty to curate its platform in ways that enable it to responsibly serve this role. In 2014, Facebook's own Community Standards introduction read: "Facebook gives people around the world the power to publish their own stories, see the world through the eyes of many other people, and connect and share wherever they go." This special position Facebook occupies in public life makes it even more imperative that categorical prohibitions be removed and a proportionality standard is introduced until Facebook develops more detailed guidance. While it is true that a proportionality standard interpreted and applied at Facebook's sole discretion is no automatic guarantee of appropriate respect for free speech, it at least introduces an evaluative step into content moderation compared to black-and-white bans.

4. A western-centric approach to human rights?

One thing that should be noted at the outset is the seemingly western-centric evolution of freedom of expression standards, at least as regards social media platforms. The majority of relevant provisions and guiding documents comes from the United Nations, the Council of Europe, and the European Union, and only marginally from the Inter-American Commission on Human Rights (part of the Organization of American States), the African Union, and the Organization for Security and Co-operation in Europe, and the Association of Southeast Asian Nations.

[195] Jürgen Habermas, Sara Lennox, and Frank Lennox, 'The Public Sphere: An Encyclopedia Article' (1964) 3 New German Critique 49.

Unsurprisingly, Facebook's own Oversight Board heavily references the instruments issued by western-centric organizations.[196]

However, the combination of the western origin of most global platforms such as Facebook, and the relative abundance of instruments and guidance on freedom of expression by western-centric organizations should not automatically be construed as that freedom of expression on Facebook reflects a purely western approach. The issue is complex and does not lend itself to a clear-cut answer,[197] nor can its intricacies be analyzed herein. We simply note here that there is evidence that the historical development of universal human rights has been influenced by non-western civilizations as well,[198] and it is reasonable to conclude that different aspects of human rights, and freedom of expression in particular, have been shaped in different degrees by western and non-western ideals.[199]

While it is likely that Facebook's understanding of freedom of expression leans toward a western conception, particularly considering that Facebook's current policies reflect also Facebook's historical western tendencies before it became a global platform, Facebook today does take a global approach. Facebook explains that its "Content Policy team, which sits in more than a dozen locations around the world, is responsible for developing [the] Community Standards and Community Guidelines" and in performing this task they "factor in cultural differences on what is acceptable and [the] different perspectives on safety and voice and the impact of [Facebook's]

---

[196] See the decisions issued by the Oversight Board at https://oversightboard.com/decision/.

[197] On the intricacies of this question see Raimundo Pannikar, 'Is the Notion of Human Rights a Western Concept?' (1982) 30 Diogenes 75.

[198] Surya Subedi, 'Are the Principles of Human Rights Western Ideas: An Analysis of the Claim of the Asian Concept of Human Rights from the Perspectives of Hinduism' (1999) 30 California Western International Law Journal 45; Janne Mende, 'Are Human Rights Western—And Why Does It Matter? A Perspective from International Political Theory' (2021)17 Journal of International Political Theory 38. In favor of human rights as a western construct: Adamantia Pollis and Peter Schwab, 'Human Rights: A Western Construct with Limited Applicability' in Christine Koggel (ed), *Moral Issues in Global Perspective – Volume 1: Moral and Political Theory* (Broadview Press, 2006) 1.

[199] Heiner Bielefeldt, '"Western" Versus "Islamic" Human Rights Conceptions? A Critique of Cultural Essentialism in the Discussion on Human Rights' (2000) 28 Political Theory 90.

policies on different communities globally".[200] Therefore, our approach herein, whereby we juxtapose Facebook's policies with various international instruments, does reflect the approach the global platform of Facebook attempts to take.

5. Social media platforms and the international community are on different speeds

International organizations and global social media companies operate quite differently, despite the apparent similarity that they both promulgate types of legal ordering regimes. Facebook's early motto was "move fast and break things,"[201] whereas international organizations operate on a consensus-based model underpinned by exhaustive negotiations, political compromises, and intricate diplomacy, which necessarily takes time. The difference in the pace of evolution of international human rights standards and of social media company policies is exacerbated when one considers that social media platforms catalyze the emergence of social phenomena (such as fake news) or magnify existing problems (such as hate speech), putting additional pressure on the competent actors to provide human rights guidance or institute new legal frameworks. As a result, as valuable as the work of international organizations can be, it is at the same time by nature, more time-intensive than the rate of events on platforms such as Facebook.

Whether the decision-making process of international institutions can be adapted to respond in a timely fashion to the needs of the accelerated digital communities that are quickly becoming the main loci where freedom of expression is exercised is an open question. The pace of evolution of international standards and guidance is necessarily dictated by the institutional constraints, inherent features and working methods of the bodies that produce them. To illustrate the relevance of these considerations it suffices to compare the instruments included in our research that have been produced at the United Nations level: aside from the treaty provisions, the guiding

[200] 'How We Update the Facebook Community Standards' (*Facebook*, 29 July 2021) https://transparency.fb.com/en-gb/policies/improving/deciding-to-change-standards/.

[201] 'Facebook, Inc' (*Wikipedia*, 2021) <https://en.wikipedia.org/wiki/Facebook,_Inc.#History.

documents we have taken into consideration come from the UN Treaty Bodies, and in particular the Human Rights Committee, and from the UN Special Procedures, which include not only reports from the Special Rapporteur on Freedom of Opinion and Expression, but also other thematic mandates such as the Special Rapporteur on the Sale and Sexual Exploitation of Children, or the Special Rapporteur on Counter-terrorism. Although the Treaty Bodies and the Special Procedures are both UN human rights monitoring procedures, there are significant differences between these mechanisms.

Treaty bodies are committees of independent experts that monitor implementation of the core international human rights treaties and, as such, their primary function is to examine state party reports to assess their compliance with their treaty obligations. Special Procedures, on the other hand, comprise either an individual (called "Special Rapporteur" or "Independent Expert") or a working group composed of five members with the mandate to report on the implementation of human rights norms in a specific thematic or geographic context.

Treaty bodies can also produce General Comments or General Recommendations, but these are "merely an attendant product aiming to give states guidance on the nature and scope of other obligations for their reports."[202] Since the adoption of General Comments is not the primary function of Treaty Bodies, it is unrealistic to expect swift guidance emanating from this institutional body. Moreover, General Comments are meant to provide interpretive guidance for the content of all the human rights provisions in the treaty. These considerations should not however diminish the value of General Comments, whose "legal analytical function […] advances the density of international understanding of the Covenant, and serves to prevent states parties from claiming that a Covenant obligation is limited to this or that area of its experience."[203] Additionally, General Comments have also acquired a policy recommendation function[204] that "can help both states and

[202] Nigel Rodley, 'United Nations Human Rights Treaty Bodies and Special Procedures of the Commission of Human Rights - Complementary or Competition?' (2003) 25 Human Rights Quarterly 882, 906.

[203] Hellen Keller and Leena Grover, 'General Comments of the Human Rights Committee and Their Legitimacy' in Helen Keller and Geir Ulfstein (eds), *UN Human Rights Treaty Bodies: Law and Legitimacy* (Cambridge University Press, 2012) 126.

[204] Ibid 124.

non-state actors determine their own plan of action on important policy issues."[205] If General Comments cannot therefore be expected to provide ad hoc and prompt guidance to all emerging social phenomena, they still perform a central function "by fleshing out the scope and content of vaguely articulated rights [in the treaty]."[206] Special Procedures, on the contrary, enjoy more flexibility and, as far as thematic mandates are concerned, "they are expected to make recommendations aimed at states generally (these are often reflected in resolutions on the subject matters of the mandates) and […] to other parts of the UN and the international community, including nongovernmental organizations."[207] In light of the generally-worded nature of the Human Rights Council resolutions establishing the Special Procedures mandates, Special Rapporteurs enjoy more discretion in determining the mandate's nature and scope,[208] and one of the key aspects of this mechanism is "to seek to have some effect and give some guidance in a short time frame."[209] This flexibility allows Special Procedures to be more responsive to emerging and urgent social issues.

Another key aspect that needs to be considered is the role that the independent experts themselves play in shaping the guiding documents that these bodies produce. As far as General Comments are concerned, there is not a specific procedure for selecting a topic: a suggestion by a Committee member might be sufficient to initiate the drafting of a General Comment.[210] If the way in which issues are prioritized is not particularly apparent, the persuasiveness of Committee

---

[205] Ibid 125.

[206] Ibid 126.

[207] Rodley (n 202) 888.

[208] Joanna Naples-Mitchell, 'Perspectives of UN Special Rapporteurs on Their Role: Inherent Tensions and Unique Contributions to Human Rights' (2011) 15 The International Journal of Human Rights 232, 234.

[209] Rodley (n 202) 907.

[210] Keller and Grover (n 203) 170; for example, General Comment 36 on the right to life was pushed for by Sir Nigel Rodley. See General Comment No. 36 on article 6 of the International Covenant on Civil and Political Rights (30 October 2018) CCPR/C/GC/36; UN HRC, 'Human Rights Committee Adopts General Comment on the Right to Life' (*United Nations*, 30 October 2018) https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=23797&LangID=E.

members and considerations about the relevance of a specific issue or the availability of sufficient practice or experience in dealing with an issue are major factors for selecting a topic.[211] The first draft of a General Comment is usually made available to the public in order to receive written submissions from state and non-state actors for comments. While the goal is to draft a General Comment that attracts the greatest possible consensus[212] (and General Comments are in fact adopted on the basis of consensus after a second paragraph-by-paragraph reading of a draft),[213] it is also true that each Committee member will bring their own legal backgrounds, interests, policy considerations and expertise to the discussion.[214] The entire process is quite lengthy, and although individual members play a significant role in shaping it, it is also the result of extensive discussions and negotiations between all the Committee members.

In contrast, the individualized nature of the UN Special Procedures has a more meaningful effect on the interpretive guidance that each mandate produces. As already mentioned, Special Rapporteurs generally enjoy discretion in interpreting the scope of their mandate and are best positioned to respond to urgent matters. The personal legal backgrounds, interests, policy considerations and expertise of mandate holders can result in either more generous or more conservative interpretations than other bodies. However, mandate holders must also operate strategic choices about how best to expend extremely limited time, human resources, and funding.[215] As such, urgency and potential impact are factors that also affect the selection of topics that a Special Rapporteur will consider during their mandate. These inherent differences might

---

[211] Keller and Grover (n 203) 170.

[212] Ibid 173.

[213] Ibid 176.

[214] Ibid 175; taking again General Comment 36 as an example, it is worth noting that during its adoption Yuval Shany, Committee Chairperson and Rapporteur for the draft General Comment, "expressed hope that General Comment no. 36 had managed to capture Sir Rodley's deep humanitarian sensibility, commitment to the legal discipline, and common sense."; See General Comment No. 36 (n 210).

[215] Naples-Mitchell (n 208) 242.

also explain why guidance produced by UN Treaty Bodies and reports produced by UN Special Rapporteurs and civil society groups is perceived as being unclear or inconsistent.[216]

Facebook is by nature more agile than international organizations. It is true that the process of revision of its content policies "involves regularly getting input from outside experts and organizations to ensure we understand the different perspectives that exist on free expression and safety, as well as the impacts of our policies on different communities globally" and the team responsible for such revisions "runs a meeting [every few weeks] to discuss potential changes to [the] policies based on new research or data" making the whole process an iterative exercise that goes through various steps and involves numerous people.[217] However, the fact that revisions across Facebook's content policies take place even at a monthly pace, shows that the company has a constant pipeline of revisions coming through to reflect latest developments, research, and decisions.

[216] Barrie Sander, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation' (2020) 43 Fordham International Law Journal 939.

[217] Mark Zuckerberg, 'A Blueprint for Content Governance and Enforcement' (*Facebook*, 5 May 2021) https://www.facebook.com/notes/751449002072082/.

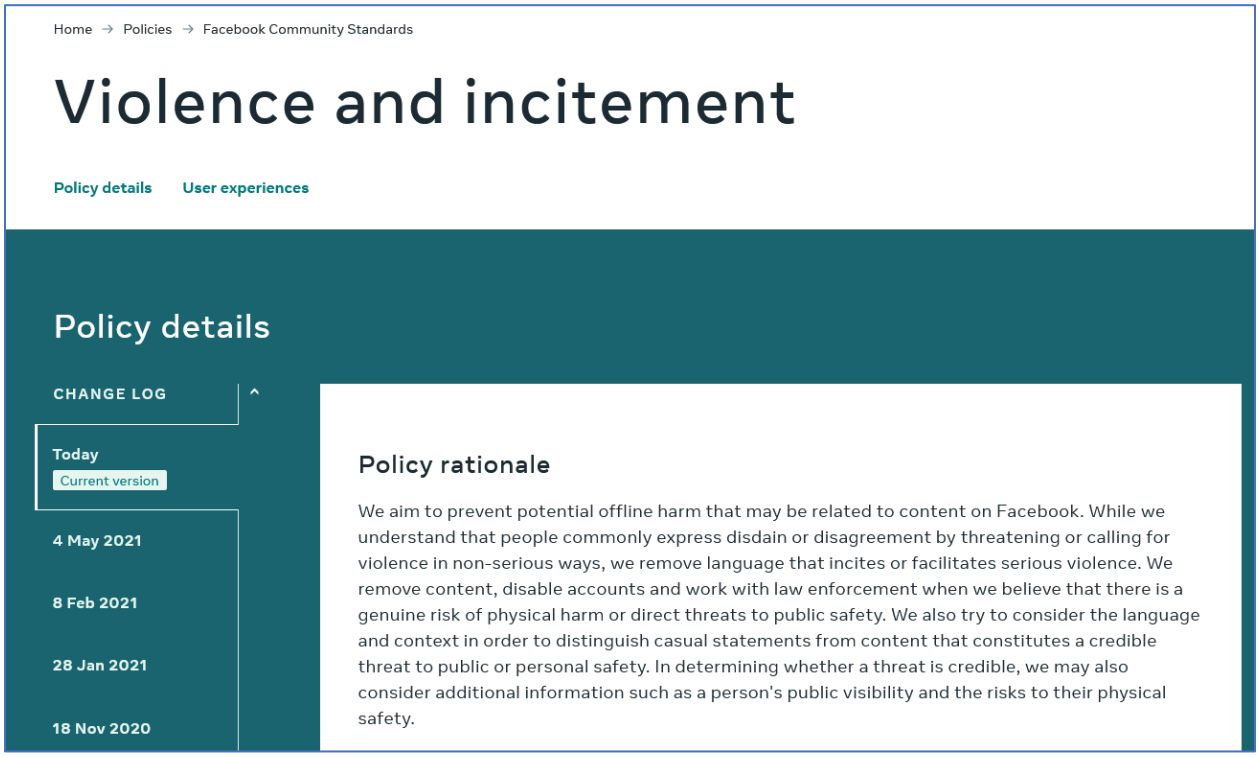


Figure 3: Screenshot of Facebook's Community Standards showing frequent revisions at a much faster pace than the norm for international instruments.

6. "It's complicated": The bright examples, the missed opportunities, and the failures of Facebook's content policies and of the international community

Social media platforms have been a long time in the making. Their precursors in the form of forums and bulletin boards existed since the Internet's commercialization in the early 90s. By the time social media networks appeared in 2003 (or even 1997 if one counts SixDegrees as the first social media network), both the concept and some of the associated freedom of expression concerns around their operation were known.

Extensive intermediary liability rules, sectoral rules that were struck down in courts (particularly on the protection of children online) on free expression grounds, as well as litigation on Nazi memorabilia, set the tone early on regarding the challenging aspects of freedom of expression online. At the very least, the problematic aspects around certain types of speech, such as

terrorism or hate speech, the role of anonymity online, the redress mechanisms users had against intermediaries, including social media networks, and the relationship between intermediaries and governments were well-identified issues either from the early days of the Internet, or at the latest around the time social media networks began to emerge.

Where this is the case, the lack of meaningful guidance on the side of the international community and the lack of balanced and detailed policies on the side of Facebook are therefore regrettable and impactful. On the other hand, one can be disappointed at slow and insufficient measures, because there have been, after all, instances where the response was more satisfactory, and it serves to raise expectations and set a desirable standard of activity. It is difficult to theorize on why certain issues have been handled better than others, but it still stands to reason that the capacity for timely and adequate rule-making is there, if the resources are committed.

a) The bright example: Tackling fake news and misinformation

Arguably the most striking example of quick and to-the-point reflexes both from the international community and from Facebook comes from the response to the rising threat of fake news. The emergence of fake news is narrowly linked to the 2016 US Presidential elections and Brexit referendum, which are often cited as examples of their disruptive impact.[218] Fake news presented a novel threat that very much affected and implicated online platforms—although not exclusively. As described by Levy, the level of fake news circulating on Facebook during the final weeks of the election campaign incremented significantly[219]. Two days after the election, Mark Zuckerberg stated that “the idea that fake news on Facebook, of which it's a very small amount

[218] McGonagle, Tarlach. ‘“Fake news” False Fears or Real Concerns?’ (2017) 35 Netherlands Quarterly of Human Rights 203.

[219] Steven Levy, *Facebook: The Inside Story* (Penguin Books, 2020).

of the content, influenced the election in any way, [...] is a pretty crazy idea."[220] The Russian interference in the 2016 US elections was discovered in the following months.[221]

Only a few months later, in March 2017, David Kaye, who at the time was UN Special Rapporteur on Freedom of Opinion and Expression, recognized the risk that "efforts to counter [fake news] could lead to censorship, the suppression of critical thinking and other approaches contrary to human rights law," and in response co-led the Joint Declaration on Freedom of Expression and Fake news. The Joint Declaration sought to identify the applicable human rights standards, to encourage the promotion of diversity and plurality in the media, and to emphasize the particular roles played by digital intermediaries as well as journalists and media outlets. While the Joint Declaration does not provide overly detailed guidance, and it does not necessarily take into account the various peculiarities of social media platforms, the very fact that it came out immediately after the fake news phenomenon exploded, and highlighted the problem, was enough of a first response. Social media platforms could no longer pretend that misinformation had not become a problem of global proportions, one that needed to be addressed, and one that placed social media platforms among the key intermediaries to be in the position to act in that direction.

With this perception of a shared problem, some social media platforms (including Facebook) and advertisers adhered to a Code of Practice on Disinformation,[222] building on the European Commission´s Communication on Tackling Online Disinformation in 2018.[223] The commitments taken by the signatories of this code include the following benchmarks with regard to content

---

[220] Casey Newton, 'Zuckerberg: The Idea that Fake News on Facebook Influenced the Election is "Crazy"' (*The Verge*, 10 November 2018) https://www.theverge.com/2016/11/10/13594558/mark-zuckerberg-election-fake-news-trump.

[221] Brian Ross, Rhonda Schwartz, and James Gordon Meek, 'Officials: Master Spy Vladimir Putin Now Directly Linked to US Hacking' (*ABC News*, 15 December 2016) https://abcnews.go.com/International/officials-master-spy-vladimir-putin-now-directly-linked/story?id=44210901.

[222] European Commission Code of Practice of Disinformation (2018) https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.

[223] Communication from the European Commission, 'Tackling Online Disinformation: a European Approach' (2018) COM/2018/236.

moderation: (1) to deploy policies and processes to disrupt advertising and monetization incentives for relevant behaviours, such as misrepresenting material information about oneself or the purpose of one's properties; (2) to keep complying with the requirement set by EU and national laws, and outlined in self-regulatory Codes, that all advertisements should be clearly distinguishable from editorial content; (3) to enable public disclosure of political advertising and use reasonable efforts towards devising approaches to publicly disclose "issue-based advertising;" (4) to put in place clear policies regarding identity and the misuse of automated bots on their service.

In September 2020, the European Commission published a first assessment of the implementation of the Code,[224] which recognizes some important achievements, and highlights areas for improvement. Among these, the need to tackle not simply "imposter websites" (i.e., sites that misrepresent their identity or purpose, or scrape content from other sources in order to generate income from ad placements) but also websites that consistently spread misinformation, the need to provide more country-level transparency, and the need to provide more tools for users to flag disinformation. The Commission followed up providing specific suggestions on how the code could be strengthened,[225] but left their implementation ultimately to the discretion of its signatories.

The handling of false news by both Facebook and the international community is a good example of quick reaction, even if Facebook has received a great deal of criticism for the actual implementation of its policies.

The handling of false news on Facebook's platform is a good example of quick reaction to a pressing problem, even if Facebook has received a great deal of criticism for the actual

[224] Staff Working Document, 'Assessment of the Code of Practice on Disinformation Achievements and Areas for Further Improvement' (2020) SWD(2020)180.

[225] European Commission, 'Guidance on Strengthening the Code of Practice on Disinformation' (2021) COM(2021)262final.

implementation of its policies.[226] False news started becoming an issue on online platforms as early as 2013, but it was not until 2016 during the US Presidential elections that false news took centerstage as a social problem and a contentious issue for online platforms.[227] On top of the media pressure, Facebook may have also taken account of the recommendations put forth by international institutions, which as mentioned previously, also showed quick reflexes in appealing to online platforms to curb the spread of misinformation.

Initially, in the very first version of its Terms of Service, Facebook simply stated that "Facebook is not responsible for any incorrect or inaccurate Content posted on the Web site or in connection with the Service..." (Terms of Service version of November 2005). For the next decade no revisions specifically concerned false news.

Unsurprisingly, considering the heat online platforms received for enabling commercial and state interests to use online platforms for election meddling purposes or public health misinformation campaigns, Facebook's Advertising Policy was the first to respond to the growing wave of misinformation, and in fact it showed a rather drastic policy change from ban to curbing. As early as 2015, Facebook's Advertising Policy banned "deceptive, false, or misleading content, including deceptive claims, offers, or business practices [and] content that exploits controversial political or social issues for commercial purpose" (version of August 2015). These restrictions are much stronger than Facebook's general Terms of Service rules around Facebook posts.

This would change in future. In October 2019 the rules narrowed considerably, only banning adverts that "include claims debunked by third-party fact-checkers, or, in certain circumstances, claims debunked by organisations with particular expertise." Until August 2019, Facebook's third-party fact checkers were appointed by the company to only vet content posted to social network by users. But a policy update published late in summer 2019 allowed fact checkers to flag false adverts for the first time. The new policy was introduced quietly by the company, and initially

---

[226] Olivia Solon, 'Facebook Failure: Did Fake News and Politarized Politics Make Trump Elected?' *The Guardian* (10 November 2016).

[227] Kate Connolly, Angelique Chrisafis, Poppy Mcpherson, Stepahnie Kirchgaessner, Benjamin Haas, Dominic Phillips, Ele Hunt, and Michael Safi, 'Fake News: An Insidious Trend That's Fast Becoming a Global Problem' *The Guardian* (02 December 2016).

noticed for the effect it had on political adverts: fact checkers are not allowed to vet content posted by political candidates, and so those adverts can never be taken down for misinformation.[228]

In April 2018, Facebook updated its Community Standards to include a section on false news. A number of influential publications, including The New York Times, had reported in 2018 that the Burmese military harnessed Facebook over several years to disseminate hate propaganda, false news and inflammatory posts—the media pressure might have contributed to Facebook's policy change.[229] Facebook's policy is not to remove false news, but rather to demote it (a more proportionate step, at least initially). The policy explains that "there is [...] a fine line between false news and satire or opinion. For these reasons, we don't remove false news from Facebook, but instead significantly reduce its distribution by showing it lower in the News Feed" (Community Standards version of November 2020). An example of such measure comes in April 2019 when Facebook introduced a new metric known as "Click-Gap" that analyzes sites and posts that generate many clicks and links on Facebook compared to the Internet as a whole. If a post seems to only be popular on Facebook and nowhere else online, then its reach will be limited in the News Feed. This update will hurt sites whose content's sole purpose is to go viral on Facebook.[230]

"While these measures are generally positive, they are an insufficient response to the challenges posed by disinformation. … content moderation efforts continue to display the same long-standing problems of inconsistent application of companies' terms of service, inadequate redress mechanisms and a lack of transparency and access to data that hampers an objective assessment of the effectiveness of the measures that have been adopted. Furthermore, although the platforms are global businesses, they do not appear to apply their policies consistently across all geographical areas or to uphold human rights in all jurisdictions to the same extent."

---

[228] Alex Hern, 'Facebook Fact Checkers Did Not Know They Could Vet Adverts' *The Guardian* (26 October 2019).

[229] Paul Mozur, 'A Genocide Incited on Facebook, With Posts from Myanmar's Military' *The New York Times* (15 October 2018).

[230] Emily Dreyfuss, 'Facebook Is Changing News Feed (Again) to Stop Fake News' (*Wired*, 18 April 2019).

UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression

In January 2020 Facebook also added a "manipulated media" provision that prohibits the posting of media that have been altered to distort their message. Reportedly, this provision was added following a video of Nancy Pelosi, slowed down to seventy-five percent speed, giving the impression that Pelosi was mentally unwell or intoxicated.[231] The policy explicitly covers only misinformation produced using AI, meaning "shallow fakes" – videos made using conventional editing tools – though frequently just as misleading, are still allowed on the platform.[232]

Even though Facebook's response to the rise of fake news was mostly successful, it should still be taken in the broader context of the exercise of the right to freedom of expression on its platform. Effective and consistent enforcement of the stated rules as well as effective remedies remain essential in the proper safeguarding of freedom of expression. As the UN Special Rapporteur for freedom of expression noted in her most recent report on Disinformation and freedom of opinion and expression, "[w]hile these measures are generally positive, they are an insufficient response to the challenges posed by disinformation. … content moderation efforts continue to display the same long-standing problems of inconsistent application of companies' terms of service, inadequate redress mechanisms and a lack of transparency and access to data that hampers an objective assessment of the effectiveness of the measures that have been adopted. Furthermore, although the platforms are global businesses, they do not appear to apply their policies consistently across all geographical areas or to uphold human rights in all jurisdictions to the same extent."[233]

---

[231] Kate Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2019) 129 Yale Law Journal 2418.

[232] Alex Hern, 'Facebook Bans "Deepfake" Videos in Run-up to US Election,' *The Guardian* (7 Jan 2020).

[233] Irene Khan, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (2021) A/HRC/47/25.

b) Attempting to get proportionality right: Bullying and harassment

Bullying and harassment, as potentially legitimate limitations to freedom of expression, is an area where Facebook's position has changed significantly over time, initially in response to the advancement of international guidance, and subsequently, on its own initiative. Bullying was recognized as a distinct category of violation of Community Standards since 2011. Importantly, the prohibition was cast by Facebook in broad terms, including not only traditional bullying but also scenarios where individuals are "being persistently contacted against their wishes" (Community Standards version of January 2011), on the premise that contacting strangers or people one has never met in person can be a form of harassment. Two points were initially controversial: the lack of definition of bullying, and the lack of definition of "private individuals" that can be targeted with this type of behavior. In 2013, Facebook provided some clarity on the rationale of the public-private distinction, asserting in connection with bullying and harassment that "users are allowed to speak freely on matters of public interest" (Community Standards version of January 2013). However, it was only in 2017, a couple of years after the international community addressed cyberbullying as part of the Report of the Special Rapporteur on Minority Rights[234], that the Community Standards provided more details on both the aforementioned issues. The updated Community Standards not only introduce a requisite intentionality for content to be degrading or shaming target, but also mention specific types of activities that would fall within the scope of the prohibition: pages that identify and shame private individuals, images altered to degrade private individuals, photos or videos of physical bullying posted to shame the victim, sharing personal information to blackmail or harass people, and repeatedly targeting other people with unwanted friend requests or messages. The Community Standards also define the notion of private individuals as people who have "neither gained news attention nor the interest of the public, by way of their actions or public profession" (Community Standards version of January 2017).

[234] Izsák (n 188).

Since 2018, Facebook stresses the importance of people's visibility also for the purpose of determining its response to violent threats, in particular to determine the credibility of threats of violence, theft, vandalism, or other financial harm. From that year, however, the concrete application of this category becomes more blurred, as the company carves out an exception for content that is "newsworthy, significant or important to the public interest" (Community Standards version of January 2018). Specifically, the Community Standards make a pledge to permit open and critical discussion of people who are featured in the news or have a large public audience based on their profession or chosen activities, while at the same time asserting that "credible threats to public figures are removed just as for private individuals." This could reasonably be taken to suggest that only a subset of bullying, the one involving credible threats, can be acted upon against public figures. This interpretation found confirmation in the latest update of the Community Standards in 2020 (version of November 2020), which announced a policy of removal of "severe attacks to public figures, or those where the public figure is directly tagged in the post or comment." At the same time, the Standards seems to endorse a layered approach, extending protection of private individuals against content meant to degrade or shame, and announcing a more restrictive approach for children, due to the risk of more serious emotional impact, prohibiting even "softer" types of offenses. As part of the same update, Facebook also pointed to the Bullying Prevention Hub, a resource that is made available for teenagers, parents and educators seeking support for issues related to bullying and other conflicts.

This latest update adds many more examples to its illustrative list of prohibited speech, including comparisons to animals or insects that are culturally perceived as intellectually or physically inferior, or to an inanimate object ("cow," "monkey," "potato"), content manipulated to highlight, circle or otherwise negatively draw attention to specific physical characteristics (nose, ear, etc.). It also adds a provision that appears to reveal a concern with the overbroad application of the prohibition, explaining that people are allowed to share and reshare posts if it is clear that something was shared in order to condemn or draw attention to bullying and harassment, and that in certain instances self-reporting is required, precisely to help understand this type of situations. The changes introduced most recently, particularly in 2018 and in 2020, are a welcome development as they fill previous gaps of protection of bullying, while also striving to safeguard freedom of expression by imbuing the norms with some form of proportionality (in the form of a layered approach) and, to prevent an overbroad application, by carving out content that is made available to actually condemn or report bullying.

c) The blind leading the blind: How terrorism became the blind spot for the international community and Facebook alike

Terrorism is a key area where social media platforms have struggled to strike the right balance, but, at the same time, despite the numerous treaties and declarations put forward over the past decades, little guidance actually existed at the international level either, at least until very recently. Without pressing and detailed guidance from the international community, Facebook was also late to adopt specific terms on terrorism and practically given carte blanche to take down content on the suspicion of incitement to terrorism.

Considering how sensitive the matter is and the considerable margin of appreciation left to states to tackle it, as well as how controversial the topic has been on platforms like Facebook (see, e.g. the 2013 beheadings controversy),[235] it is only natural to expect that anti-terrorism measures would clash with freedom of expression. Yet, it was only in 2011 that the friction between anti-terrorism measures and freedom of expression was elevated to a key consideration. In General Comment No. 34 the UN Human Rights Committee underscored the vagueness that surrounds anti-terrorism measures, and called for "such offences as 'encouragement of terrorism' and 'extremist activity' as well as offences of 'praising', 'glorifying', or 'justifying' terrorism [to] be clearly defined to ensure that they do not lead to unnecessary or disproportionate interference with freedom of expression. Excessive restrictions on access to information must also be avoided. The media plays a crucial role in informing the public about acts of terrorism and its capacity to operate should not be unduly restricted. In this regard, journalists should not be penalized for carrying out their legitimate activities."[236]

---

[235] Leo Kelion, 'Facebook Lets Beheading Clips Return to Social Network' (*BBC News*, 21 October 2013) https://www.bbc.co.uk/news/technology-24608499.

[236] General Comment No. 34 (n 193).

It is evident, that despite the numerous instances where the conflict between freedom of expression and anti-terrorism legislation was identified, it all came too late, too vaguely, and without much consideration to the peculiarities of the online social networking environment.

A vague warning in the 2007 Guidelines of the Committee of Ministers of the Council of Europe on Protecting Freedom of Expression and Information in Times of Crisis that "member states should not use vague terms when imposing restrictions of freedom of expression and information in times of crisis"[237] had preceded General Comment 34, and the same was repeated again in the subsequent Joint Declaration on Freedom of Expression and Responses to Conflict Situations adopted in 2015 which highlighted the need for States to "refrain from applying restrictions relating to 'terrorism' in an unduly broad manner. Criminal responsibility for expression relating to terrorism should be limited to those who incite others to terrorism; vague concepts such as 'glorifying', 'justifying' or 'encouraging' terrorism should not be used."[238] Expressing concerns that, in light of the absence of a clearly agreed definition of "terrorism" in international law, states had a broad margin of discretionary power to interpret what kinds of expression constitute incitement to terrorism, the Report also drew attention to the definition suggested by the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, which requires "(a) an intent to incite the commission of a terrorist offence; and (b) the existence of an actual risk that such an offence will be committed as a consequence."[239] Similarly, the Report underscored that "any domestic criminal laws that prohibit incitement to terrorism must meet the three-part test of restrictions to the right to freedom of expression," which means that "incitement of terrorism: (a) must be limited to the incitement of conduct that is truly terrorist in nature, as properly defined; (b) must restrict the right to freedom of expression no more than is

[237] Committee of Ministers of the Council of Europe, 'Guidelines on Protecting Freedom of Expression and Information in Times of Crisis' (2007) CM/Del/Dec(2007)1005/5.3.

[238] 'Joint Declaration on Freedom of Expression and Responses to Conflict Situations' (*United Nations*, 2015) https://www.ohchr.org/en/NewsEvents/Pages/DisplayNews.aspx?NewsID=15921&LangID=E.

[239] Ibid.

necessary for the protection of national security, public order and safety or public health or morals; (c) must be prescribed in law in precise language, including by avoiding reference to vague terms such as 'glorifying' or 'promoting' terrorism; (d) must include an actual (objective) risk that the act incited will be committed; (e) should expressly refer to two elements of intent, namely intent to communicate a message and intent that this message incite the commission of a terrorist act; and (f) should preserve the application of legal defences or principles leading to the exclusion of criminal liability by referring to "unlawful" incitement to terrorism."[240]

In addition, the Joint Declaration on Freedom of Expression and Countering Violent Extremism adopted in 2016 stressed that "everyone has the right to seek, receive and impart information and ideas of all kinds, especially on matters of public concern, including issues relating to violence and terrorism, as well as to comment on and criticise the manner in which States and politicians respond to these phenomena" and that the concepts of "violent extremism" and "extremism" should not be used as the basis for restricting freedom of expression unless they are defined clearly and appropriately narrowly.[241]

It is evident, that despite the numerous instances where the conflict between freedom of expression and anti-terrorism legislation was identified, it all came too late, too vaguely, and without much consideration to the peculiarities of the online social networking environment. The newly-passed Regulation on addressing the dissemination of terrorist content online[242] includes a host of new measures that online intermediaries should take to curb the spread of terrorist content on their platforms.[243] These new measures are not yet in effect, but they are expected to significantly contribute to how platforms like Facebook handle terrorist content.

---

[240] Fionnuala Ni Aoláin, 'Promotion and Protection of Human Rights and Fundamental Freedoms while Countering Terrorism' (2021) A/76/261.

[241] Dunja Mijatovic, 'Communiqué by the OSCE Representative on Freedom of the Media on the Impact of Laws Countering Extremism on Freedom of Expression And Freedom of the Media' (2014) Communiqué N. 6/2014; Dunja Mijatovic, 'Communiqué by the OSCE Representative on Freedom of the Media on Free Expression and the Fight Against Terrorism' (2016) Communiqué No. 6/2016.

[242] Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online.

[243] Part 1 Section 4.d.

Facebook introduced the prohibition of organizations with a record of "terrorist or violent criminal activity" from maintaining a presence on their website only in 2013 (almost ten years after it started offering its services). The company did not however provide a definition of "terrorist activity," notwithstanding the existing criticism precisely on the fact that lack of definitional clarity and the discretionary application of rules can result in undue restriction of the right to freedom of expression that was already advanced in 2011 by the Special Rapporteur on Freedom of Opinion and Expression.

Facebook only updated its policy on terrorism and provided a clear definition in 2020 (see below). The delay was despite constant criticism against social media companies for their role in serving as platforms for dissemination of terrorist content.[244] This is not to say that platforms like Facebook did not act on terrorist content—in fact, as of February 2016 dedicated teams at Facebook were proactively removing all posts or profiles with links to terrorist activity following the terrorist attacks in Paris and Brussels in late 2015.[245] But the actual Terms of Service or Community Standards, which create the binding constitution between Facebook and its community failed to reflect the backstage moderation activity.

In May 2019 Nick Clegg, Vice-President of Global Affairs and Communications at Facebook, joined G7 governments for a meeting in Paris on how to curb the spread of terrorism and extremism online, at which they signed up to the Christchurch Call to Action.[246] The technology companies also committed to a nine-point plan that sets out concrete steps the industry will take to address the abuse of technology to spread terrorist content. As an online content sharing

---

[244] Larry Greenemeier, 'Social Media's Stepped-Up Crackdown on Terrorists Still Falls Short' (*Scientific American*, 24 July 2018) https://www.scientificamerican.com/article/social-medias-stepped-up-crackdown-on-terrorists-still-falls-short/; Laurence Bindner and Raphael Gluck, 'Trends in Islamic State's Online Propaganda: Shorter Longevity, Wider Dissemination of Content' (ICCT, 5 December 2018) https://icct.nl/publication/trends-in-islamic-states-online-propaganda-shorter-longevity-wider-dissemination-of-content/.

[245] Natalie Andrews and Deepa Seetharaman, 'Facebook Steps Up Efforts Against Terrorism' *The Wall Street Journal* (11 February 2016).

[246] Jacinda Ardern, 'Christchurch Call to Action Summit' (2019) https://www.christchurchcall.com/christchurch-call.pdf.

service provider, one of the five individual actions it committed to concerned the Terms of Use: "We commit to updating our terms of use, community standards, codes of conduct, and acceptable use policies to expressly prohibit the distribution of terrorist and violent extremist content."[247] In that same year, the Special Rapporteur on terrorism criticized the Facebook definition of terrorism, defining it as an "overly broad and imprecise definition of terrorism […], which equates all non-State groups that use violence in pursuit of any goals or ends to terrorist entities."[248]

The Christchurch Call to Action commitment and the criticism moved by the UN Special Rapporteur perhaps influenced the drafting of the 2020 update to the "dangerous individuals and organizations" policy, which now includes a qualification of terrorist groups which seems to be taken from the International Convention for the Suppression of the Financing of Terrorism. However, the definition given by Facebook *includes* the international definition, and it is unclear whether the company enjoys discretion in widening its scope. It is also important to underscore that contextual analysis is still fundamental when assessing whether a particular type of content constitutes incitement to terrorism. As highlighted by the Special Rapporteur "States must ensure that their measures to address the threats of terrorism, violent extremism and protect national security do not negatively affect civil society. In particular: (a) Definitions of terrorism and of violent extremism in national laws must not be overly broad and vague. They must be precise and sufficiently narrow to not include members of civil society or non-violent acts carried out in the exercise of fundamental freedoms. Emergency measures must be strictly limited and not used to crack down on civil society actors; (b) Legitimate expression of opinions or thought must never be criminalized. Non-violent forms of dissent are at the core of freedom of expression. Reporting on, documenting or publishing information about terrorist acts or counter-terrorism measures are essential aspects of transparency and accountability. The key role of the Internet, particularly within repressive societies or for marginalized groups, must be recognized and protected."[249]

---

[247] Ibid.

[248] Ibid.

[249] Fionnuala Ní Aoláin, 'Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism' (2019) A/HRC/40/52.

The Standard mentions that Facebook removes content that refers to listed categories without context that condemns or neutrally discusses said content. While this could be a measure in compliance with protection of legitimate expression, it is also important to underscore, as also mentioned in a recent decision by the Facebook Oversight Board,[250] that the current policy does not offer "clear examples that explain the application of 'support,' 'praise' and 'representation,' making it difficult for users to understand this Community Standard" and "fails to explain how it ascertains a user's intent, making it hard for users to foresee how and when the policy will apply and conduct themselves accordingly."[251]

Although, as identified earlier, the international community failed to offer detailed guidance that took into account the peculiarities of the online environment, Facebook was unnecessarily late in addressing longstanding issues only in 2020.

d) Stricter than necessary: Facebook's approach to anonymity

As recognized by the 2013 Report of the OAS Special Rapporteur on Freedom of Expression, participation in public debate without revealing one's identity is a normal practice in modern democracies: it is conducive to the participation of individuals in public debate since—by not revealing their identity—they can avoid being subject to unfair retaliation for the exercise of a fundamental right. It does not solely entail writing opinion articles or participating in debate forums—it also involves the ability to call for social mobilizations, to call upon other citizens to protest, to organize politically, or to challenge the authorities even in risky situations.[252] Anonymity, therefore, has been an integral component of freedom of expression.

---

[250] Ibid.

[251] Facebook Oversight Board, Case Decision 2020-005-FB-UA https://oversightboard.com/news/141077647749726-oversight-board-overturns-facebook-decision-case-2020-005-fb-ua/.

[252] Catalina Botero, 'Annual Report of the Office of the Special Rapporteur for Freedom of Expression' (2013) OEA/Ser.L/V/II.149. Doc. 50.

The right to anonymity was not explicitly addressed by international standards until the 2013 Report of the UN Special Rapporteur on Freedom of Expression, despite the fact that anonymity was widely recognized early on as a key feature of Internet communications,[253] and was referred by the same Report as "one of the most important advances enabled by the Internet."[254] However, the Recommendations laid out in the report, including one to refrain from requiring the verification of identity as a precondition for access to communications, did not include prescriptions for private sector actors. The same can be said about the 2013 Report of the OAS Special Rapporteur on Freedom of Expression, which encouraged states to promote online spaces where people's activities and identities are not observed or documented, including through the preservation of anonymous platforms for the exchange of content and use of proportionate authentication services, and linked this to the State's obligation to create a safe environment for the exercise of freedom of expression.[255] This State-focused approach was replicated in the 2016 Report of the OAS Special Rapporteur, despite restating the fundamental importance for freedom of expression of preserving privacy (a point vocally made by Frank LaRue in his 2013 Report),[256] which it defined as "every personal and anonymous space that is free from intimidation or retaliation, and necessary for an individual to be able to freely form an opinion and express his or her ideas as well as to seek and receive information, without being forced to identify him or herself or reveal his or her beliefs and convictions or the sources he or she consults."[257] Specifically with regard to anonymity, it emphasized States' obligation to respect anonymous discourse as an exercise of privacy and freedom of expression, with a possibility only exceptionally to require proof of identity from the person expressing it, following a proportionality test.[258]

---

[253] Lawrence Lessig, 'The Law of the Horse: What Cyberlaw Might Teach' 113 Harvard Law Review 501 (1999).

[254] Botero (n 252) para 23.

[255] Ibid.

[256] Frank La Rue, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (2013) A/HRC/23/40.

[257] Edison Lanza, 'Standards for a Free, Open and inclusive internet' (2017) OEA/Ser.L/V/II, para 185.

[258] Ibid para 228.

The turning point in the extension of Recommendations to corporations was the 2015 Report of the UN Special Rapporteur on Freedom of Expression, which reminded them of their obligations to respect human rights independently of whether the State meets its own obligations[259] and concluded that companies, like States, should refrain from blocking or limiting the transmission of encrypted communications and permit anonymous communication.[260] While this is an important step forward in providing guidance, however, it lacks depth in discussing the very concept of anonymity: a common feature in all these documents is that they fail to distinguish between the different notions of citizen anonymity (against the State), customer anonymity (against the service provider) and platform anonymity (against the community of user on a particular platform), thus leaving significant leeway for the actors involved.[261]

Anonymity has been a contentious issue for Facebook, as the platform maintained a "real name" policy since its founding. On top of requiring the use of authentic birth name, it prohibited all misrepresentation about oneself or their affiliation, and expanding to cover misrepresentation about age in 2006. This rendered the right to speak anonymously impractical on the platform. In 2011, Facebook further specified this rule by prohibiting the creation of multiple accounts, on the grounds that this would undermine the trust and safety that accurate representation of users creates; and in 2013, the language was amended to include "creating a false presence for an organization" (a concept which could in principle be broader than misrepresenting that organization).

However, in 2017 the Community Standards embraced the possibility that users create a presence on Facebook for a pet, organization, favorite movie, games character or other purposes using a Facebook page, rather than a profile. The terms also clarify that Facebook may ask page owners to associate their name and Facebook Profile with a Page that contains cruel and insensitive content (a category that is specifically defined elsewhere), which seems to imply that in normal circumstances a page does not need to be associated with a user´s real name.

---

[259] David Kaye, 'Report on Encryption, Anonymity, and the Human Rights Framework' (2015) A/HRC/29/32, para 28

[260] Ibid para 62.

[261] Nicolo Zingales, 'Virtues and Perils of Anonymity: Should Intermediaries Bear the Burden?' (2014) 5 JIPITEC 155, paras 6-7.

While one can be sympathetic to the rationale for the prohibition against "inauthentic" or duplicate profiles, the Facebook real name policy's clash with the exercise of fundamental rights cannot be denied, even more so after the recommendations made in this sense by several rapporteurs on Freedom of Expression.

Another opening to the pseudo-/anonymous use of Facebook is the reference in the policy regarding the *possibility* of closing an account where it is discovered that a user has multiple personal profiles, which could be taken to imply that such sanction does not apply to all cases. In fact, this appears to be in line with Facebook's positioning with regard to people who use different names from the one they were born with, including transgender people and victims of domestic violence who use aliases to hide from their abusers. In 2014, Facebook made a public announcement that they would build new authentication tools to verify accounts for people who use in real life names that are different from their birth names. This happened when a coalition led by drag queens in San Francisco pressured the company to review its system, after several performers reported problems with the "real name" policy.[262] However, in 2015 it was reported that the updated policy (which gives several options for verification, including non-official documents like library cards and fidelity cards[263]) still made it difficult for people who had recently changed their name and had no documents ready for verification.[264]

No significant changes were made in this respect in the updates of the Community Standards from 2018 and 2020. By contrast, in 2020, authenticity was elevated to "the cornerstone of our

[262] Casey Newton, 'Facebook Clarifies Real Name Policy Amid LGBT Protests' (*The Verge*, 1 October 2014) https://www.theverge.com/2014/10/1/6881641/facebook-will-update-real-name-policy-to-accommodate-lgbt-community.

[263] Facebook, 'What types of ID does Facebook accept?' (*Facebook*, 2021) https://www.facebook.com/help/159096464162185.

[264] Sam Levin, 'As Facebook Blocks the Names of Trans Users and Drag Queens, this Burlesque Performer is Fighting Back' *The Guardian* (29 June 2017).

community," explicitly including amongst violations the creation of "inauthentic profiles" and the following acts as well: creating another Facebook or Instagram account after being banned from the site; creating or managing a Page, group, event or Instagram profile because the previous Page, group, event or Instagram profile was removed from the site; evading the registration requirements outlined in [the] Terms of Service Impersonate others by Using their images with the explicit aim to deceive people; and creating a profile assuming the persona of or speaking for another person or entity.

All in all, while one can be sympathetic to the rationale for the prohibition against "inauthentic" or duplicate profiles, its clash with the exercise of the fundamental right of freedom of expression cannot be denied, even more so after the recommendations made in this sense by several Rapporteurs. The issue has been only partly addressed by Facebook through the use of its discretion in the non-enforcement of the provision in specific cases,[265] and by opening up the possibility of using a broader range of documents for verification purposes. Unfortunately, this does not address the vast impact that a verification policy can have on legitimate uses of the social network – including the engagement in socially valuable research, as discussed below in the section on intellectual property and access to knowledge.

e) (Not) giving users the means to challenge authority: Facebook's lacklustre remedies and redress policies and the international community's late mobilization.

Users who object to how their activity on social media platforms is treated would benefit from remedies and redress mechanisms, meaning ways that they can challenge the decisions social media platforms take against them. Access to effective remedies is also one of the areas Ruggie

[265] Justin Osofsky, 'Comunity Support FYI: Improving the Names Process on Facebook' (*Facebook*, 15 December 2015) https://about.fb.com/news/2015/12/community-support-fyi-improving-the-names-process-on-facebook/.

identified as imperative to be guaranteed both by States and corporations;[266] as a result, meaningful guidance from the international community toward social media companies is most welcome. Certainly, some guidance has been provided, although one could perhaps quarrel over the sufficiency of its depth and scope, or its timeliness in fleshing out platform-specific context.[267] For instance, the right to an effective remedy is enshrined in the Universal Declaration of Human Rights of 1948 (art. 8) as well as its transposition into regional human rights convention (1953 in Europe, 1969 in America and 1981 in Africa), the International Covenant on Civil and Political Rights of 1966 (art. 2) and the Convention on the Elimination of Racial Discrimination of 1965 (art. 6); however, it is formulated in a generic sense in these provisions. It can also be found in a more context-specific formulation in the Convention on the Rights of the Child of 1989 (art. 39), which prescribes the need to "promote physical and psychological recovery and social reintegration of victims [...] in an environment which fosters the health, self-respect and dignity." Although some general guidance directed to States on the implementation of this principle could be found in the Vienna Declaration and Programme of Action in 1993, and, with regard to judicial mechanisms, in General Comment 13 on the administration of justice, it was only in 2005 that more detailed guidance was produced, when the UN General Assembly adopted and proclaimed the Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law.[268] While this guidance refers to a specific set of circumstances - those involving particularly serious violation of human rights - the framework set out there provides a helpful reference to operationalize the right in other scenarios too, by breaking it down into three components: (a) Equal and effective access to justice; (b) Adequate, effective and prompt reparation for harm suffered; and (c) Access to relevant information concerning violations and reparation mechanisms. The Basic Principles also elaborate on the

---

[266] Guiding Principles on Business and Human Rights (n 173).

[267] In 2019 OHCHR launched the B-Tech Project, which aims to provide authoritative guidance and resources for implementing the United Nations Guiding Principles on Business and Human rights (UNGPs) in the technology space, including in the area of Accountability and Remedy https://www.ohchr.org/EN/Issues/Business/Pages/B-TechProject.aspx.

[268] UNGA, Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law, 21 March 2006, A/RES/60/147.

forms under which effective reparation should take place: restitution, compensation, rehabilitation, satisfaction and guarantees of non-repetition.[269]

However, this guidance did not address the responsibilities of non-State actors, which were only recognized in 2011 with the adoption by the UNGPs.[270] The UNGPs explicitly call business enterprises to "establish or participate in effective operational-level grievance mechanisms for individuals and communities who may be adversely impacted" (principle 29) and lay out several criteria to ensure the effectiveness of non-judicial grievance mechanisms (principle 31), listed below:

- Legitimate, enabling trust from stakeholder groups for whose they are intended enabling trust from the stakeholder groups for whose use they are intended, and being accountable for the fair conduct of grievance processes;
- Accessible: being known to all stakeholder groups for whose use they are intended, and providing adequate assistance for those who may face particular barriers to access;
- Predictable: providing a clear and known procedure with an indicative time frame for each stage, and clarity on the types of process and outcome available and means of monitoring implementation;
- Equitable: seeking to ensure that aggrieved parties have reasonable access to sources of information, advice and expertise necessary to engage in a grievance process on fair, informed and respectful terms;
- Transparent: keeping parties to a grievance informed about its progress, and providing sufficient information about the mechanism's performance to build confidence in its effectiveness and meet any public interest at stake.
- Rights-compatible: ensuring that outcomes and remedies accord with internationally recognized human rights;
- A source of continuous learning: drawing on relevant measures to identify lessons for improving the mechanism and preventing future grievances and harms;

[269] Ibid paras 18-23.

[270] Guiding Principles on Business and Human Rights (n 173).

- Based on engagement and dialogue: consulting the stakeholder groups for whose use they are intended on their design and performance, and focusing on dialogue as the means to address and resolve grievances.

Importantly, the commentary on principle 31 (h) clarifies that "dialogue" implies that a business enterprise cannot both be the subject of complaints and unilaterally determine the outcome; and that where adjudication is needed, this should be provided by a legitimate, independent third-party mechanism.

More recent documents provide a further layer of contextualization, focusing on the specific realities of online activity. For instance, in 2017, in its thematic report on digital access providers, the UN Special Rapporteur on Freedom of Expression zoomed in on digital access providers.[271] In parallel, the Council of Europe Recommendation on Intermediaries in 2018 addressed specifically Internet intermediaries, requiring them to ensure human review of automated content management "where appropriate" and to provide accessible, equitable, expedient complaint mechanisms that compatible with rights, transparent and affordable, and with built-in safeguards to avoid conflicts of interest when the company is directly administering the mechanism.[272] It also established minimum standards applicable across the board as a requisite for effective remedies: namely, the existence of an impartial and independent review of the alleged violation, which may result in inquiry, explanation, reply, correction, apology, deletion, reconnection or compensation.[273]

More guidance on platforms' remediation procedures is provided in the 2018 Report on Online Content Regulation by the UN Special Rapporteur on Freedom of Expression: after reminding at the outset that States bear the primary duty to remediate business-related human rights abuses (especially those that they instigate) and that companies may cause or contribute to such abuses if they fail to provide for or cooperate in their remediation through legitimate processes, the Report

[271] David Kaye, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (2017) A/HRC/35/22, paras 74-75.

[272] Council of Europe, Recommendation of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries (2018) CM/Rec(2018)2.

[273] Ibid.

points to specific instantiations of such remediation process, ranging from reinstatement and acknowledgment to settlements related to reputational or other harms.[274] To this, one should add the specific recommendations provided by the Rapporteur in its 2018 thematic Report on artificial intelligence, where it prescribed that businesses inform individuals that they have been subject to automated decisions, equip them with information about the logic behind that decision, ensure the human review of requests for remedy, and publish data on the frequency at which remedial mechanisms are triggered.[275] It also acknowledges the potential of more innovative and cost-effective remedy solutions such as user flagging and company-specific or industry-wide ombudsman programmes and social media councils, in particular as it would allow to hear individual users' complaints that meet certain criteria and gather public feedback on recurrent content moderation problems such as over-censorship related to a particular subject area.[276] However, the Rapporteur also warned that if the failure to remediate persists, legislative and judicial intervention may be required.[277]

The Rapporteur continued this thread in his 2019 Report on State Surveillance, where he clarified that the State duty to provide an effective remedy implies not only that law enforcement and prosecutorial authorities should investigate allegations of violations promptly, thoroughly and effectively through independent and impartial bodies, but also an obligation to protect individuals against acts by private sector entities that cause infringements, by exercising due diligence to prevent, punish, investigate or redress the harm caused by such acts by private persons or entities.[278] More importantly, for our purposes, it stated that private companies engaged in surveillance must at a minimum develop a number of safeguards, including: notification processes that promptly report misuses of their tools to the relevant government oversight bodies (such as national human rights institutions) or intergovernmental bodies (such as special procedures

[274] David Kaye, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (2019) A/74/486, para 59.

[275] David Kaye, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (2018) A/73/348 para 60.

[276] Ibid para 58.

[277] Ibid.

[278] Kaye (n 188).

complaints mechanisms); grievance mechanisms that enable individuals to submit complaints concerning human rights abuses facilitated by company products and services, and provide for independent assessment of those complaints and meaningful follow-up[279] ; and remedial mechanisms that enable complainants to seek compensation, apologies and other forms of redress, as appropriate, in cases in which complaints are independently verified.[280]

In the same year, the Rapporteur laid out important work on the subject for the 2019 Annual Report, devoted to online hate speech, where it held that the companies should at a minimum publicly identify the kinds of remedies that they will impose on those who have violated their hate speech policies, and should have graduated responses according to the severity of the violation or the recidivism of the user.[281] Going more at length into the type of approaches that can be taken, the Report acknowledges the role of de-amplification and de-monetization of problematic expressions, the potential insufficiency of user suspension (but also the need for post-violation impact assessment) and the promises of remedial policies such as apology, education, public denunciation, counter-speech, reporting and training.[282]

In addition to these context-specific documents, the UN produced detailed cross-sectoral guidance in 2020, with the report issued by the Office of the High Commissioner for Human Rights on its Accountability and Remedy Project.[283] This is undoubtedly the most advanced set of recommendations for the operationalization of the right to effective remedies, and it will be interesting to see how those recommendations will feed into the practice of stakeholders (in particular business enterprises, which are addressed in Part III[284]). Particularly noteworthy is the

---

[279] Ibid.

[280] Ibid para 60.

[281] Kaye (n 274).

[282] Ibid para 55.

[283] Office of the High Commissioner for Human Rights, Accountability and Remedy Project: Improving Accountability and Access to Remedy in Cases of Business Involvement in Human Rights Abuses, https://www.ohchr.org/EN/Issues/Business/Pages/OHCHRaccountabilityandremedyproject.aspx.

[284] Human Rights Council, Improving Accountability and Access to Remedy for Victims of Business-Related Human Rights Abuse Through Non-State-Based Grievance Mechanisms, Report of the United Nations High Commissioner for Human Rights' (2020) A/HRC/44/32.

explicit reference to the need to cater to people who may be at heightened risk of vulnerability or marginalization (policy objective 7.4), as well to address power imbalances, conflict of interests and undue influence (policy objective 7.6).

Providing an effective remedy is a crucial component of businesses' fulfillment of their human rights due diligence, and it cannot be denied that Facebook has embraced at least some of this responsibility by materially improving the status quo over the last twelve years.

Providing an effective remedy is a crucial component of businesses' fulfillment of their human rights due diligence, and it cannot be denied that Facebook has embraced at least some of this responsibility by materially improving the *status quo* over the last twelve years. However, it also cannot be said that Facebook´s *status quo* has typically been in tune with international standards, due to the numerous shortcomings that have been observed throughout the process. All in all, one could describe the evolution of Facebook´s remedial policy as a process of imperfect, but incremental improvement. To start, remedies available to users for potential violations of their rights did not appear high priorities in Facebook´s policies until 2009, when the opportunity to appeal against wrongful takedown for copyright was introduced in its Terms of Service (version of June 2009). One can criticize the narrow focus, as the opportunity to appeal was not explicitly provided against all possible legally wrongful takedowns, regardless of the type of violation at stake (a generalized right to request a review was only recognized in 2019), and more generally, on the grounds that Facebook did not provide "easy-to-use mechanisms to report conduct or content," as recommended in 2009 by the Safer Social Networking Principles for the EU (Section 4), or "easily accessible information on how to report and complain about interferences with your rights and how to seek redress," as demanded by the CoE's Guide to Human Rights for Internet Users in 2014.

The latter concern was addressed in 2011 (Community Standards version of January 2011), when Facebook introduced a general right to report anything that users see on the site and that they believe that violates its terms. It was accompanied by the cautionary disclaimer that this does not guarantee removal from the site (but if the content is in violation of the Community Standards,

it may be removed and subject in some cases to “legal or other action”), and by the reminder that Facebook offers “personal controls over what you see, such as the ability to hide or quietly cut ties with people, Pages, or applications that offend you”.

In 2012, Facebook introduced in its Terms of Service (version of June 2012) a mention to removal as a remedial action for trademark infringement, in particular in relation to complaints made about users’ choice of name or other identifiers for account and pages. Then, starting in 2017, Facebook’s Community Standards (version of January 2017) began to speak more broadly about removals, partly anticipating the guidance given on remedies by the Special Rapporteur on Freedom of Expression later that year in its report on digital access providers. The Community Standards clarified that the following content will be removed, which presumably includes both proactive and reactive content moderation, that is, upon complaint: credible threats of physical harms to individuals, hate speech, specific threats of theft, vandalism, or other financial harm (including hate speech and credible threats made against public figures), content that expresses support for groups that are involved in violent or criminal behavior, content that appears to purposefully target private individuals with the intention of degrading or shaming them, content that threatens or promotes sexual violence or exploitation, photographs or videos depicting incidents of sexual violence and images shared in revenge or without permissions from the people in the images, descriptions of sexual acts that go into vivid detail, and photographs of people displaying genitals or focusing in on fully exposed buttocks. It is also acknowledged that in some cases the remedy against a violation may be a restriction, rather than a removal, as it is the case for some images of female breasts if they include the nipple, or with regard to distasteful or offensive content that users can avoid by using “certain Facebook tools” (presumably, the need to consent before viewing). Additionally, the terms mention Facebook’s collaboration with law enforcement when there is a genuine risk of physical harm or direct threats to public safety, and the virtues of counter-speech, and the availability of tools for counter-speech (in the form of accurate information and alternative viewpoints) and communication with the person who posted disagreeable or disturbing content in order to resolve possible issues. The entirety of remedial tools made available against violations is further clarified in the 2018 update of Facebook’s Terms of Service (version of June 2018), making reference to offering help, removing content, blocking access to certain features, disabling an account, and contacting law enforcement.

The 2017 update (version of January 2017) provides also more information about the reporting process, clarifying that review decisions may occasionally change after receiving additional

context about specific posts or after seeing new violating content appearing on a Page or Facebook Profile, and that the consequences for a violation depend on the severity of the violation and the person's history on Facebook, but not on the number of reports received in relation to that piece of content. Broadly speaking, this is an alignment with the proportionality principle, and at the same time a positioning against the possible abuse of reporting tools, which is also recognized since 2020 as a violation of Facebook's authenticity policy.

In 2019 (Terms of Service version of September 2019), on occasion of the introduction of a generalized right to request a review, along with a commitment to let users know and explain any options they have for review, the terms carved out a few exceptions: namely, when the user has seriously or repeatedly violated the terms; when doing so may expose Facebook or others to legal liability; when this "harm[s] our community of users;" when it compromises or interferes with the integrity or operation of any of Facebook's services, systems or Products; and where Facebook is restricted due to technical limitations or prohibited from doing so for legal reasons. This seems to curtail the effectiveness of the right to a remedy, leaving significant discretion to Facebook in granting the opportunity to request a review. Furthermore, it should be noted that the mere existence of a review may not be sufficient to comply with procedural justice principles, in particular as this does not necessarily involve consideration of the user's arguments in the way that an appeal does. Perhaps more strikingly, this review does not involve a right to receive an explanation for the removal, which strikes as contrary to the concept of "notice" and "right to be heard" that are recognized as part of international due process standards[285] and were explicitly recalled by the 2014 CoE Guide on Human Rights for Internet Users and the 2017 and 2018 Reports of the UN Special Rapporteur on Freedom of Expression, among others.

In 2020, the latest update of Facebook´s terms (Community Standards November 2020) included references to a number of additional removal possibilities (imagery believed to be in violation of a user's privacy rights; bullying and harassment on a memorialized profile, nude images of children; images that depict incidents of sexual violence and intimate images shared without permission from the people pictured; copyright or trademark infringing material following receipt of a report from a rights holder or an authorized representative; and content that displays,

[285] Charles T. Kotuby, Jr. and Luke A. Sobota, General Principles of Law and International Due Process: Principles and Norms Applicable in Transnational Disputes (Oxford University Press, 2017).

advocates for or coordinates sexual acts with non-consenting parties or commercial sexual services, such as prostitution and escort services) and a couple of additional informational resources, such as a Bullying Prevention Hub (a resource offering step-by-step guidance, including information on how to start important conversations about bullying) and a guide to reporting and removing intimate images shared without one's consent. It is interesting to note that with this update Facebook moves beyond the requirements imposed by international standards in the specific domains of concern, broadly recognizing the key importance of human rights due diligence across the board. This allowed the company to provide solutions in areas that are not yet addressed by international law, as is the case for instance for the so-called "revenge pornography" (i.e., the non-consensual sharing of intimate images/videos) and the handling of memorialized profiles, and to follow the recommendation of the 2020 Report of the Special Rapporteur on the Sale and Sexual Exploitation of Children, including child prostitution, child pornography and other child sexual abuse material, that ICT serve as an essential element of successful prevention and response strategies, supporting efforts of law enforcement agencies and non-governmental organizations (affirming that "where domestic laws have not yet caught up with international standards, private sector stakeholders have an opportunity to bring their practices in line with these standards and promote innovative solutions and positive change.")[286]

Furthermore, we would be remiss not to mention what constitutes arguably the most innovative attempt by Facebook to implement individuals' right to remedy against adverse decisions taken by its content moderators, the creation of the Oversight Board and the commitment taken in its Bylaws and reflected by Facebook in its Community Standards (since 2021) to implement the Board's content decisions unless doing so could violate the law. It should be noted, however, that the small number of cases that can be decided upon by the Board and the vagueness of the values that should guide its interpretation of the Community Standards dramatically reduces the effectiveness of this remedy.[287] Furthermore, as noted by former UN Special Rapporteur David

[286] Boer-Buquicchio (2020) (n 188) para 33.

[287] Kate Klonick (n 231); Evelyn Douek, 'Facebook's "Oversight Board": Move Fast with Stable Infrastructure and Humility' (2019) North Carolina Journal of Law & Technology 1, Stefania Di Stefano, 'The Facebook Oversight Board and the UN Guiding Principles on Business and Human Rights: A Missed

Kaye,[288] the Board falls short of the international standards for an effective remedy in several respects: first of all, the Board is not required to apply international human rights law, but only to ”pay particular attention” to it. Secondly, it is not empowered to order remedies beyond reinstatement, which in itself may be insufficient to restore justice[289]. Third, it does not provide sufficient transparency over its work, and in particular the process that leads to the selection of cases. Fourth, and perhaps most importantly, the Board is not fully independent from Facebook.[290]

Oversight Board aside, another important observation concerns the fact that the framework provided by Facebook’s Community Standards does not detail the way in which the principle of proportionality is taken into account in the selection and adoption of remedies. This is a particularly important matter in the context of online platforms, on the one hand due to the shift to a user-generated content environment, and on the other hand due to the more diverse set of responses that algorithmic technologies enable content moderators to take, compared to broader array of tools than simply prohibiting publication, removing, extending the right of reply or compensating for possible harms: for instance, “flagging” or “modulating views.”[291] Eric Goldman has provided a comprehensive taxonomy of content moderation remedies, distinguishing among: (1) actions against individual content items; (2) actions against an online account; (3) actions to reduce the visibility of violations, which can be implemented against individual content items or an entire account; (4) actions to impose financial consequences for violations, which also can be

---

Opportunity for Alignment?’ in Jonathan Andrew and Frédéric Bernard (eds) *Human Rights Responsibilities in the Digital Age – States, Companies and Individuals* (Hart Publishing, 2021).

[288] See Research Report by the Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, with the support of the International Justice Clinic at the University of California, Irvine School of Law (2020) https://ohchr.org/documents/issues/opinion/researchpaper2020.pdf.

[289] Ibid paras 42-60.

[290] Ibid paras 60-67.

[291] Ivar Hartmann, ‘A New Framework for Content Moderation’ (2020) Computer and Security Law Review 36.

implemented against individual content items or an entire account, and (5) a miscellaneous category for other actions that do not fit into the prior four categories[292] (See

Figure 4). Goldman also discusses the factors that ought to be weighted in the determination of the appropriate type of remedy, including (1) severity of the rule violation; (2) confidence that a rule violation actually occurred; (3) scalability and consistency; (4) the community's ability to self-correct; (5) how the remedies impact others; (6) retaining user engagement while curbing violations and recidivism; and (7) the potential application of parallel sanctions, both judicially and extrajudicially.[293]

Content Regulation	Account Regulation	Visibility Reductions (by acct or item)	Monetary (by acct or item)	Other
• Remove content • Suspend content • Relocate content • Edit/redact content • Interstitial warning • Add warning legend • Add counterspeech • Disable comments	• Terminate account • Suspend account • Suspend posting rights • Remove credibility badges • Reduced service levels (data, speed, etc.) • Shaming	• Shadowban • Remove from external search index • Nofollow authors' links • Remove from internal search index • Downgrade internal search visibility • No auto-suggest • No/reduced internal promotion • No/reduced navigation links • Reduced virality • Age-gate • Display only to logged-in readers	• Forfeit accrued earnings • Terminate future earning (by item or account) • Suspend future earning (by item or account) • Fine author/impose liquidated damages	• Educate users • Assign strikes/warnings • Outing/unmasking • Report to law enforcement • Put user/content on blocklist • Community service • "Restorative justice"/apology

Figure 4: Content Moderation Remedies (Source: Goldman, 2022).

[292] Eric Goldman, 'Content Moderation Remedies' Michigan Technology Law Review (Forthcoming 2022).

[293] Ibid.

The debate about remedy proportionality has been front and center in the recent decision by Facebook's Oversight Board on the indefinite suspension of the account of Donald Trump, where the Board asked Facebook to determine and justify a proportionate response that is consistent with the rules that are applied to other users of its platform.[294] Among other criticisms, the Board took issue with the lack of criteria to define the suspension period, the lack of a clear, published procedure on suspensions, and the lack of guidance regarding the application of content moderation decisions to users with large audiences. When discussing the proportionality of the suspension, the Board referred to the factors listed in the Rabat Plan of Action to assess capacity of speech to create a serious risk of inciting discrimination, violence, or other lawless action: context, status of the speaker, intent, content and form, extent and reach of the speech, and imminence of harm. On this basis, the Board concluded that the initial suspension for 24 hours and its subsequent extension were necessary and proportionate measures to prevent severe human rights harm, but the application of this sanction for an indefinite period was not. Interestingly, the Decision also features a diverging opinion from the minority of the Board, urging the proportionality analysis to be informed by Mr. Trump's use of Facebook's platforms prior to the November 2020 presidential election, in line with the Rabat Plan of Action's[295] consideration of the frequency, quantity, and extent of harmful communications to determine the level of incitement.

Overall, the assessment of Facebook's implementation of remedial mechanisms is mixed: one should combine the positive performance in meeting international standards on certain aspects of the right to a remedy with the shortcoming in the functioning of remedial mechanisms, in particular the lack of explanation and effective contestation against the application of remedies that generate adverse effects on users. Moreover, one could criticize Facebook's slow uptake of the recommendations that were provided by the international community since 2008, and even

---

[294] Case 2021-001-FB-FBR *Facebook, Donald Trump v. Facebook* (2021).

[295] Human Rights Council, 'Annual Report of the United Nations High Commissioner for Human Rights, Addendum: Report of the United Nations High Commissioner for Human Rights on the Expert Workshops on the Prohibition of Incitement to National, Racial or Religious Hatred' (2018) A/HRC/22/17/Add.4.

earlier if one considers that State-focused recommendations could have been relied upon as benchmarks in the construction of private dispute resolution mechanisms.

f) Better late than never: The long-winded road to detailed guidance on hate speech

Hate speech has always been a remarkably problematic area not only for social media platforms but also for states, and drawing the line between what constitutes a legitimate exercise of the right to freedom of expression and what constitutes incitement to hatred is not an easy task.

International law lacks a clear definition on hate speech, although the EU Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law offers a definition that is also adopted by the Code of Conduct on Countering Illegal Hate Speech Online. This definition, however, has been criticized for being overbroad and incompatible with international standards on freedom of expression;[296] as such, it is "likely to create more legal uncertainty for users and, most worryingly, lead to the application of the lowest common denominator when it comes to the definition of 'hate speech.'"[297] Guidance tailored to the manifestations of hate speech on social media platforms was not provided until recently. However, international law has provided the tools, further refined during the years, for defining the boundaries of what acceptable speech is and what speech is to be prohibited, which could have guided the development of content moderation policies even in the absence of specific guidance for social media platforms.

Article 20(2) of the ICCPR expressly prohibits "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence," but does not offer much guidance as to how to draw the boundaries of this category of speech.

---

[296] Article 19, The European Commission's Code of Conduct for Countering Illegal Hate Speech Online (2016).

[297] Ibid para 19.

The growing concerns about the rise of racism and xenophobia online prompted the Council of Europe to adopt, already in 2003, the Additional Protocol to the Convention on Cybercrime. As highlighted in its explanatory report, both the Convention on Cybercrime and its Additional Protocol were drafted precisely as a response to the emergence of the Internet, which provides a new platform to "certain persons with modern and powerful means to support racism and xenophobia and enables them to disseminate easily and widely expressions containing such ideas."[298] This instrument is devoted to the criminalization of acts of a racist and xenophobic nature committed through computer systems and provides definitions for categories of content that State parties are obliged to criminalise under their domestic jurisdictions.

In 2011, the UN Special Rapporteur on the Promotion and Protection of the Rights to Freedom of Opinion and Expression released a report focusing specifically on the right to freedom of expression on the Internet.[299] Referring explicitly to the fact that "the dissemination of 'hate speech' via the Internet has also spurred efforts to regulate online content,"[300] he further clarified that advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence is one of the exceptional types of expression that States are required to prohibit under international law.[301] He acknowledged the lack of a definition of hate speech in international law and noted that many forms of hate speech do not meet the level of seriousness set out in Article 20(2) of the ICCPR, but he also referred to the recently adopted General Comment 34 which clarifies the relationship and complementarity of Article 19 of the ICCPR, which protects freedom of expression, and art. 20, which prohibits incitement to discrimination, hostility, or violence. [302]

---

[298] Council of Europe, 'Explanatory Report to the Additional Protocol to the Convention on Cybercrime, Concerning the Criminalisation of Acts of a Racist And Xenophobic Nature Committed Through Computer Systems' (2003), para 3.

[299] Frank La Rue, 'Promotion and Protection of the Right to Freedom of Opinion and Expression' (2011) A/66/290.

[300] Ibid para 26.

[301] Ibid para 28.

[302] Ibid para 26-27.

> ***"The companies should define how they determine when a user has violated the hate speech rules. At the present time, it is difficult to know the circumstances under which the rules may be violated. There seems to be very significant inconsistency in the enforcement of rules. The opacity of enforcement is part of the problem. A set of factors is identified in the Rabat Plan of Action that is applicable to the criminalization of incitement under article 20 (2) of the Covenant, but those factors should have weight in the context of company actions against speech as well. They need not be applied in the same way as they would be applied in a criminal context. However, they offer a valuable framework for examining when the specifically defined content – the posts or the words or images that comprise the post – merits a restriction".***
>
> ***UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression***

The concerns about the need to further clarify the relationship between these two provisions resulted in the Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence,[303] the outcome document of a four-year initiative led by OHCHR, which was presented in 2013. Although the guidance offered by the Rabat Plan of Action is addressed primarily to states, this document provides much needed clarifications on how content moderation policies can deal with hate speech. It outlines a six-part threshold test for defining restrictions on freedom of expression, incitement to hatred, and for the application of article 20 of the ICCPR, taking into account (1) the social and political context, (2) status of the speaker, (3) intent to incite the audience against a target group, (4) content and form of the speech, (5) extent of its dissemination and (6) likelihood of harm, including imminence.[304]

The 2019 Report of the UN Special Rapporteur on Freedom of Opinion and Expression, which is devoted to online hate speech, explains the relevance of the Rabat Plan for platforms such as

[303] Addendum on the Prohibition of Incitement to National, Racial or Religious Hatred (n 295).

[304] Ibid Appendix para 29.

Facebooks as follows: “The companies should define how they determine when a user has violated the hate speech rules. At the present time, it is difficult to know the circumstances under which the rules may be violated. There seems to be very significant inconsistency in the enforcement of rules. The opacity of enforcement is part of the problem. A set of factors is identified in the Rabat Plan of Action that is applicable to the criminalization of incitement under article 20 (2) of the Covenant, but those factors should have weight in the context of company actions against speech as well. They need not be applied in the same way as they would be applied in a criminal context. However, they offer a valuable framework for examining when the specifically defined content – the posts or the words or images that comprise the post – merits a restriction.”[305]

Other UN Special Procedures addressed hate speech in their reports, including the Special Rapporteur on Minority Issues,[306] the Special Rapporteur on freedom of religion or belief[307] and the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance.[308]

While it is true that detailed guidance for social media platforms specifically was released relatively late, the explicit prohibition of any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence in the ICCPR and the adoption of the Additional Protocol to the Convention on Cybercrime already in 2006, which precisely targets racism and xenophobia committed via the Internet, signify an early response to an issue that is challenging, ever evolving and particularly complex. Moreover, the regular attention that the international community has shown to the issue throughout the years is an indication not only of

[305] Kaye (n 274) para 49.

[306] Izsák (n 188).

[307] Heiner Bielefeldt, ‘Report of the Special Rapporteur on Freedom of Religion or Belief’ (2015) A/HRC/31/18; Ahmed Shaheed, ‘Report of the Special Rapporteur on freedom of Religion or Belief’ (2019) A/HRC/40/58.

[308] E. Tendayi Achiume, ‘Report of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance’ (2020) A/HRC/44/57.

the continuous relevance of the topic, but also of the need to constantly monitor a phenomenon that is widespread, severe and multifaceted.

Facebook already introduced a prohibition of "hateful speech" in the 2005 Terms of Service, but the alignment of its hate speech policies with international standards has been rather slow. Facebook's hate speech policies attracted public attention in 2018, following the publication of the Report of the UN Fact-finding Mission in Myanmar, which underscored the role that Facebook played in the incitement of violence in the country, stating that "[t]he role of social media is significant. Facebook has been a useful instrument for those seeking to spread hate, in a context where, for most users, Facebook is the Internet. Although improved in recent months, the response of Facebook has been slow and ineffective."[309]

When introduced in 2005, the prohibition of "hateful speech" lacked a clear definition, notwithstanding the fact that pre-2005 international instruments such as Article 20(2) of the ICCPR, which expressly prohibits "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence," already provided a baseline for defining hate speech (although, as mentioned earlier, without detailing how to draw clear boundaries between legitimate speech and hate speech). Similarly, the Additional Protocol to the Convention on Cybercrime, which addresses specifically the criminalization of acts of a racist and xenophobic nature committed through computer systems, also offered some definitions that could have been borrowed by the company. However, given that at this time Facebook's user base was not particularly diversified and that the UNGPs had not been adopted yet, it might not have been unreasonable for Facebook not to draw from international law provisions directly already in 2005.

In 2013 Facebook updated its Community Standards on hate speech following the controversy around the "Innocence of Muslims" video (posted on YouTube),[310] with those aggrieved alleging that the video fostered anti-Muslim sentiment. Klonick explains that Facebook grappled with this

---

[309] Human Rights Council, 'Report of the Independent International Fact-Finding Mission on Myanmar' (2018) A/HRC/39/64.

[310] 'The Anti-Islam-Film Riots: A Timeline' (*The Week*, 18 September 2012) http://theweek.com/articles/472285/antiislamfilm-riots-timeline, as cited in Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 Harvard Law Review 1598, 1624.

hate speech issue with a single rule: attacks on institutions (countries, religions or leaders was deemed to be permissible content, whereas attacks on groups (people of a certain religion, race or country) would be taken down.[311] In the same year, the Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence is adopted by OHCHR: the Rabat Plan of Action introduces the requirement for contextual analysis when assessing whether a particular type of constitutes a legitimate exercise of the right to freedom of expression or incitement to hatred. While Facebook introduced here an element of contextual analysis, it is still significantly vaguer than the requirements set out in the Rabat Plan of Action.

The inadequacy of the approach to contextual analysis persisted in the following years: for example, Facebook's 2017 Direct Threats policy, while constituting an attempt in complying with the Rabat Plan of Action, contained a presumption around troubled regions which is prejudicial and therefore problematic. This presumption was removed in 2018 and "violent and unstable regions" was replaced by "likelihood of real-world violence." This latter definition, however, could be either over-inclusive or under-inclusive, although it represents a step forward from assuming credibility if the threat is against people living in violent and unstable regions. While the introduction of this element, coupled with considerations on the public visibility of the speaker, is a hint towards greater contextual analysis, the criteria against which the analysis is carried out are not defined, and the "person's public visibility" and "likelihood of real-world violence" do not fully account for the criteria in the Rabat Plan of Action. In particular, the Rabat Plan of Action requires an assessment of the likelihood (including imminence) of discrimination, hostility or violence actually occurring: by limiting such an occurrence only to real-world instances, this definition excludes all forms of harm that are less tangible, or do not amount to *physical* harm to individuals but could still constitute hostility or discrimination – a user could feel threatened even when the harm occurs only online.

The most substantial updates to the hate speech policies were introduced in 2020, with some of them bringing the Community Standards more in line with those set in international human rights law. With the introduction of tiers, which indicate the level of content enforcement and

---

[311] Klonick ibid 1624, citing Jeffrey Rosen, 'The Delete Squad', (*New Republic*, 29 April 2013) https://newrepublic.com/ article/113045/free-speech-internet-silicon-valley-making-rules.

outline content subcategories, the Community Standards provide far more granular examples of what constitutes hate speech with regard to particular minorities.

Antisemitism has also been an area that has attracted criticism. In 2019, the Special Rapporteur on Freedom of Religion or Belief expressed serious concerns about the increased frequency of antisemitic incidents and the prevalence of antisemitic hate speech online, and underscored that "social media companies should take reports of cyberhate seriously, enforce terms of service and community rules that do not allow the dissemination of hate messages, provide more transparency in their efforts to combat cyberhate and offer user-friendly mechanisms and procedures for reporting and addressing hateful content" and "[t]hey should also report criminal antisemitic behaviour online to relevant local law enforcement agencies, including expression that constitutes incitement to discrimination, hostility or violence."[312] In August 2020, conspiracy theories about Jewish people "controlling the world" were explicitly banned from Facebook and Instagram for the first time.[313] The decision came after criticism of slow response to British grime artist Wiley's antisemitic posts.[314] It is only in October 2020 that Facebook updated its hate speech policy to prohibit any content that denies of distorts the Holocaust. Explaining the shift in policy, Zuckerberg said "my own thinking has evolved as I've seen data showing an increase in antisemitic violence, as have our wider policies on hate speech."[315] With this update to the hate speech policy, which prohibits any content that denies or distorts the Holocaust, Facebook aligned its Community Standards with international standards on freedom of expression and freedom of religion.

Lastly, while political affiliation is still not listed as a protected characteristic in the Community Standards, Facebook has now included in the list of prohibited speech "content targeting a person or group of people on the basis of their protected characteristic(s) with [...] Political exclusion,

---

[312] Ahmed Shaheed 'Report of the Special Rapporteur on Freedom of Religion or Belief' (2019) A/74/358, para 87.

[313] Alex Hern, 'Facebook and Instagram Ban Antisemitic Conspiracy Theories And Blackface' *The Guardian* (12 August 2020).

[314] Ibid.

[315] James Clayton, 'Facebook Bans Denial Holocaust Content' (*BBC News*, 12 October 2020) https://www.bbc.com/news/technology-54509975.

defined as denial of right to political participation," bringing the standard in line also with Article 7 of the Convention on All Forms of Discrimination Against Women, which provides that "State Parties shall take all appropriate measures to eliminate discrimination against women in the political and public life of the country and, in particular, shall ensure to women, on equal terms with men, the right …[t]o vote in all elections and public referenda and to be eligible for election to all publicly elected bodies".

If the latest hate speech policies offer more granular examples of what constitutes hate speech, contextual analysis remains a central element when assessing whether a specific piece of content rises to the level of incitement to discrimination, hostility or violence.

If the latest hate speech policies offer more granular examples of what constitutes hate speech, contextual analysis remains a central element when assessing whether a specific piece of content rises to the level of incitement to discrimination, hostility or violence. Although our analysis only considers Facebook's content policies until 2020, we acknowledge here that the 2021 Community Standard on Hate Speech has introduced a set of criteria for evaluating whether a piece of content constitutes a threat of harm, which include but *are not limited to* "content that could incite imminent violence or intimidation; whether there is a period of heightened tension such as an election or ongoing conflict; and whether there is a recent history of violence against the targeted protected group." The Community Standard also acknowledge that, *in some cases,* Facebook "may also consider whether the speaker is a public figure or occupies a position of authority." These elements bring the Community Standard closer to the approach set out in the Rabat Plan of Action, which, as mentioned above, requires to consider (1) context; (2) speaker; (3) intent; (4) content and form; (5) extent of the speech act; (6) likelihood, including imminence. However, Facebook should state more clearly all factors taken into account to assess the context in which the speech occurs and should give more attention to the speaker in question, who should not be taken into account only in some, undefined, instances. The contextual analysis is necessary for every category of content listed in these standards if freedom of expression is to be protected.

g) The perils of categorical bans: Facebook's unjustifiable nudity policy

Nudity has been a thorny issue for Facebook and unnecessarily so. There is no reason why the detailed policies Facebook started developing in 2013, almost ten years after it began offering its services, could not have been implemented earlier, particularly against the backdrop of the gold standard of legality, necessity, and proportionality, which was already well established by the international community, as well as more specialized instruments that concerned child pornography (Convention on the Protection of Children Against Sexual Exploitation and Sexual Abuse (2007)), which could be used as a source of inspiration on definition of pornographic content. Surely, Facebook aims to limit explicit content that goes beyond pornographic content, and this is a reasonable goal, but the blanket prohibition was bound to be problematic.

Under Article 19(3) of the ICCPR, freedom of expression can be restricted for protecting the rights of others or for protecting public morals. These restrictions, as already mentioned, must be prescribed by law and necessary and proportionate to pursue a legitimate aim. The concept of public morals is difficult to define. In General Comment 22, the Human Rights Committee stated that "the concept of morals derives from many social, philosophical and religious traditions; consequently, limitations on the freedom to manifest a religion or belief for the purpose of protecting morals must be based on principles not deriving exclusively from a single tradition."[316] However, these limitations are also to be interpreted in light of the universality of human rights and the principle of non-discrimination.[317] With respect to the protection of the rights to others, a restriction on nudity could be justified on the grounds of protecting the privacy of victims of non-consensual intimate image sharing (Article 17 ICCPR), and the rights of the child to life and development (Article 6, CRC), which are threatened in cases of sexual exploitation (as also underscored by the Oversight Board in a recent decision[318]). Nonetheless, a complete ban on

[316] Human Rights Committee, 'CCPR General Comment No. 22: Article 18 (Freedom of Thought, Conscience or Religion)' (1993) CCPR/C/21/Rev.1/Add.4.

[317] Ibid.

[318] Facebook Oversight Board, Case Decision 2020-004-IG- UA.

nudity results in a disproportionate impact on the protection of other rights, including the right to health and the right to artistic freedom.

Facebook was silent on nudity and pornography during its first five years of operation, introducing a ban in its Terms of Service on content that is "pornographic, or that contains nudity" only in 2009. The rules do not become more detailed until Facebook develops Community Standards, and even then, one has to wait until 2013 for the first details to emerge in the form of exceptions for "content of personal importance, whether those are photos of a sculpture like Michelangelo's David or family photos of a child breastfeeding." As early as 2009 pressures started mounting on Facebook to make its nudity policy more flexible, when the so-called "lactivists" held a protest outside Facebook's offices, and 11,000 mothers staged a virtual "nurse-in" online.[319] Perhaps the allusion to the famous David statue was inserted in light of the much-publicized claim brought by a French citizen against Facebook in 2011 for moderating a painting of a nude woman, L'Origine du Monde by Courbet.[320]

The gradual nuance added to Facebook's policy on nudity was welcome, but unnecessarily delayed.

Between 2016 and 2020 Facebook's Community Standards on nudity become increasingly more detailed, recognizing exceptions for "reasons like awareness campaigns or artistic projects […] photos of women actively engaged in breastfeeding or showing breasts with post-mastectomy scarring. […] photographs of paintings, sculptures, and other art that depicts nude figures" (version of March 1, 2017). The allusions to campaigning or educational purposes comes soon after the "Napalm Girl" incident involving high profile journalists and politicians in Norway in 2016.

[319] Levy (n 219) 326; Barrie Sander (n 216).

[320] Sarah Cascone, 'After an Eight-Year Legal Battle, Facebook Ends Its Dispute with a French School Teacher Who Posted Courbet's Origin of the World' (*Artnet*, 05 August 2019) https://news.artnet.com/art-world/facebook-courbet-lawsuit-ends-1616752.

Sheryl Sandberg, Facebook's COO, termed it "an iconic image of historical importance," requiring a derogation from the nudity standard. By 2020, Facebook has developed a lengthy and detailed list of "Do's and Don'ts" on nudity in response to repeated criticism of its overbroad policies that muffled freedom of speech contrary to widely accepted standards[321](e.g., April 11, 2020 version: "Do not post: Images of real nude adults, where nudity is defined as visible genitalia except in the context of birth giving and after-birth moments or health-related situations (for example, gender confirmation surgery, examination for cancer or disease prevention/assessment); visible anus and/or fully nude close-ups of buttocks unless photoshopped on a public figure …"), and has implemented age controls to show different content to adults and to children based on nudity type and levels (e.g., "We only show this content to individuals 18 and older: Real world art that depicts sexual activity; posting photographs or videos of objects that depict sexual activity in real world art; implied sexual activity in advertisements …"). Nonetheless, given that the Community Standards explicitly restrict "uncovered *female* nipples," they still raise concerns with respect to the principle of non-discrimination: this standard, coupled with the reliance on inaccurate automation for enforcement, results in disproportionate impact on women, with serious consequences not only on women's right to freedom of expression, but potentially also on other fundamental rights, including the right to health.[322]

The gradual nuance added to Facebook's policy on nudity was welcome, but unnecessarily delayed. Even when restrictions on nudity are fair to protect children and the rights of others (including privacy), any such limits still have to abide by the requirements of legality, necessity and proportionality, which have been developed at the international level for decades.[323] As Facebook's own Oversight Board has recognized, complete or overbroad bans such as those used by Facebook until very recently do not conform with legality as they are not sufficiently clear, precise and publicly accessible, nor with necessity and proportionality, because there are less intrusive means that can achieve the protective function of the limitations.[324] Moreover, as

---

[321] Danielle Keats Citron, 'Extremist Speech, Compelled Conformity, and Censorship Creep' (2018) 93 Notre Dame Law Review 1035.

[322] Facebook Oversight Board, Case Decision 2020-004-IG- UA.

[323] Ibid para 8.3.

[324] Ibid. See also General Comment 34 (n 193) paras 25, 34.

Facebook's platform was becoming more technically advanced allowing for far greater and individualized control of content flow, filters and tools that can demote or keep certain types of speech away from certain categories of people (whether because of their characteristics, e.g. children, or because they choose to), could also play a positive role, something that began to take place in 2018.[325]

h) The devil is in the details: Protected characteristics as an example of how detailed guidance can safeguard both free speech and the rights of others

Facebook introduced the concept of 'protected characteristics' only in 2011 as part of its non-discrimination policy. The principle of non-discrimination is a well-established principle under international human rights law, reiterated in every core human rights instrument. Additionally, the UN Convention on the Elimination of All Forms of Racial Discrimination (1965) the UN Convention on the Elimination of all Forms of Discrimination Against Women (1979) and the UN Convention on the Rights of People with Disabilities (2006) provide further and more tailored definitions on discrimination based on race, gender and disability respectively.

Facebook could have already prohibited discrimination on the basis of the characteristics listed in the international instruments, such as race, colour, sex,

[325] In November 2018, in a statement from Mark Zuckerberg, Facebook will be demoting content that comes close, or "borderline," to the policy line of prohibited content. For example, a post that may contain offensive speech but does not fall under hate speech will be demoted in distribution. The same goes for sexually suggestive images or ones that may tease nudity without fully showing it. This change should also affect posts that are spreading or promoting misinformation, including across the political spectrum. See Wallaro, ´Facebook News Feed Algorithm History´ (29 April 2021) <https://wallaroomedia.com/facebook-newsfeed-algorithm-history/>

language, religion, political or other opinion, national or social origin, property, birth or other status.

As such, Facebook could have integrated this principle in the content policies already in their early days, notwithstanding the fact that the UNGPs, which recognize the responsibility of business enterprises to respect human rights, were adopted only in 2011. In particular, Facebook could have already prohibited discrimination on the basis of the characteristics listed in the international instruments, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.

Instead, the early versions of Facebook's Terms of Service (version of November 2005) only make a vague reference to "racially, ethnically or otherwise objectionable" content, which seems to suggest that the company was aware of the potential exploitation of the platform for spreading discriminatory speech but did not fully flesh out a policy consistent with recognized forms of discrimination.

Interestingly, when Facebook introduced its "protected characteristics" safeguards, it also included "disability" as one of the discrimination grounds, thus demonstrating up-to-date compliance with international standards (the UN Convention on the Rights of People with Disabilities was adopted in 2006). However, while the Convention "adopts a broad categorization of persons with disabilities and reaffirms that all persons with all types of disabilities must enjoy all human rights and fundamental freedoms," Facebook introduces, in 2017, a distinction between serious and non-serious disability. This is problematic not only because it is unclear how to draw the distinction, but it also narrows the range of protection afforded by the UN Convention.

i) Chilling effects on free speech: The lack of sufficient safeguards around government requests for takedown or access to user data

Guidance has been lacking for too long in the key area of transparency in the relationship between online intermediaries, including social media platforms, and governments. As early as 2013, the UN Special Rapporteur on Freedom of Expression noted in its annual report that

domestic laws generally provide states with *carte blanche* access to communications data with little oversight or regulation.[326] However, despite making specific requests to states for the publication of statistics of communication surveillance techniques, the Report shied away from making such requests to private companies, despite the fact that this type of transparency reporting was an established practice for leading technology companies (Facebook and Google) since 2013. The OAS Special Rapporteur on Freedom of Expression merely recognized as "good practice" the publication of transparency report including at least the numbers and types of requests that lead to restrictions of freedom of expression and privacy.[327]

It was, once again, UN Special Rapporteur on Freedom of Expression David Kaye who moved the needle beyond states in his Report in 2016, when he called on social network services to be transparent in disclosing governmental requests for content takedowns.[328] He then continued on this trajectory by adding significant detail on the notion of transparency reporting in the 2017 Report, focused on the human rights roles and responsibilities in the Internet access industry: first, he prescribed the disclosure to the maximum extent allowed by law of information about government activities that require corporate assistance or involvement,[329] recommending also to adopt innovative transparency measures, such as the publication of aggregate data and the selective withholding of information, to mitigate the impact of gag orders and other non-disclosure laws.[330] Second, he required such reporting to be regular and ongoing, and in an accessible format that provides appropriate context, in order to allow civil society to challenge human rights abuses and facilitate accountability for such practices.[331]

---

[326] La Rue (n 256) para 61.

[327] Botero (n 181) para 161.

[328] David Kaye, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (2016) A/HRC/32/38.

[329] Ibid.

[330] Ibid para 71.

[331] Ibid para 70.

More recently, in his 2019 Report,[332] he endorsed the minimum standards of accountability for surveillance demanded by civil society, including among other a requirement of transparency reporting that discloses the potential uses and capabilities of one's products and the types of after-sales support provided, incidents of misuse and data concerning the number and type of sales to law enforcement, intelligence or other government agencies or their agents. The standards also feature other periodical mechanisms of transparency such as (a) regular programmes of audits and human rights verification processes to ensure that use of their products and services comply with international human rights law, including a commitment to publicly disclose key findings from these audits and verification processes; (b) notification processes that promptly report misuses of their tools to the relevant government oversight bodies (such as national human rights institutions) or intergovernmental bodies (such as special procedures complaints mechanisms); and (c) regular consultations with affected rightsholders, civil society groups and digital rights organizations about the ongoing or potential impacts of their products and services and the human rights safeguards required to prevent or mitigate these impacts, with particular emphasis on engaging those at risk of surveillance-based discrimination or repression, such as racial and ethnic minorities and historically marginalized groups.

Unsurprisingly, due to the slow advancement of international guidance on transparency with regard to government relationships, there is virtually no mention in the Community Standards of the handling of government requests. The only point in the Standards which touches on this aspect is the section on "Reporting Abuse", where Facebook recognizes that it may make unavailable in one specific country or territory, after a careful review, content that violates local law but not Facebook's Terms and Standards. No further indication is given on its relationship with government requests or the way in which these principles would be practically implemented, although it is relevant to mention that Facebook published since December 2013 periodic transparency reports covering government request for user data and restrictions based on local law,[333] and committed to continue doing so through its adherence since 2018 to the Global

---

[332] Kaye (n 188).

[333] Colin Stretch, 'Global Government Requests Report' (*Facebook*, 27 August 2013) https://about.fb.com/news/2013/08/global-government-requests-report/; Eleni Kosta and Magdalena Brewczyńska, 'Government Access to User Data: Towards more Meaningful Transparency Reports' in

Network Initiative, a multistakeholder platform created with the aim to address the challenges of protecting freedom of expression and privacy by global ICT companies (ICTs)[334]. Transparency reports are made available in Facebook's Transparency Center, together with reports on the enforcement of community standards (since January 2019), enforcement of Intellectual Property (since December 2017) and intentional Internet disruptions (since December 2016).

Facebook therefore does make available information on how it handles government requests. However, these details are not included in the legally binding Terms of Service and associated documents, and therefore do not provide users with the full legal backing a binding contract between them and Facebook would provide.

j) Take it from the international community: Facebook's late arrival at detailed rules on free speech and the protection of minors

One of the legitimate limitations to freedom of expression is the protection of minors from harmful speech. This includes both communication toward minors but also speech that involves minors (e.g. pictures). In the context of the online environment where communication with children becomes easier and more anonymous, striking the right balance between freedom of expression and protection of minors becomes both more imperative and more difficult.

Our research indicates that the international community has provided sufficient and timely guidance promptly taking into account the specific developments of the online environment. In fact, the international community started to closely monitor these issues and their manifestations in the online sphere already in the early 2000s. Regulatory efforts to address child pornography had become a major concern owing to the fact that the Internet was facilitating the distribution of this kind of content, which is explicitly prohibited under international law, and in particular by the

---

Rosa Ballardini, Petri Kuoppamäki, and Olli Pitkänen (eds), *Regulating Industrial Internet Through IPR, Data Protection and Competition Law* (Kluwer Law International, 2019) 253-274.

[334] 'Facebook joins the Global Network Initiative' (*Global Network Initiative*, 22 March 2013) https://globalnetworkinitiative.org/facebook-joins-the-global-network-initiative/.

Optional Protocol to the Convention on the Rights of the Child on the sale of children, child prostitution and child pornography.

The CoE Convention on Cybercrime, adopted in 2001, already sought to strengthen protective measures for children against sexual exploitation by criminalizing various aspects of the production, possession and distribution of child pornography when committed via the Internet and computer systems. These growing concerns around child sexual abuse and exploitation on the internet were confirmed also by the media: in 2007, for example, the New York Times reported that a "concerned parent" had opened a Facebook profile posing as a 15-year-old girl and, after signing up for three dozen sexually themed group (which were prohibited under the company's Terms of Service), received many messages from other adult users containing nude pictures and sexual advances[335].

The Lanzarote Convention, adopted in 2007, was the first international legal instrument to require the criminalization of the solicitation of children for sexual purposes (grooming), a practice that is particularly facilitated by information and communication technologies. In that direction, the Convention specifically recognizes ICTs as a category of relevant stakeholders that should participate in the elaboration and implementation of policies to prevent sexual abuse and exploitation of children, either through self-regulation or co-regulation, which demonstrates an early recognition of the significant role of the internet and social media platforms in facilitating child predation and pornography but also of their potential for ensuring their prevention and reporting.

The international community has therefore provided a timely and sufficient response to the issue of child sexual abuse and child pornography online: not only regulatory frameworks have been provided in a timely manner, but the guidance provided has been consistent and constant, and has evolved considering the specific developments of these issues and attempting to readapt regulatory frameworks to address newly emerging concerns that were specifically linked to the online environment.

[335] Brad Stone, 'New Scrutiny for Facebook Over Predators' *New York Times* (30 July 2007).

At the UN level, the Special Rapporteurs on the sale of children, child prostitution and child pornography have also produced reports that have focused on the manifestations of these conducts on the Internet, outlining how the online component creates new challenges to the international protection framework. These reports also highlighted the relevance of corporate social responsibility in this area, encouraging "the business sector to develop applications for mobile devices which allow children to report cases of online sexual abuse and exploitation, and to ensure that applications do not facilitate the sexual exploitation of children" (2014)[336], and, "[w]here domestic laws have not yet caught up with international standards, […] to bring their practices in line with these standards and promote innovative solutions and positive change" (2020)[337].

The international community has therefore provided a timely and sufficient response to the issue of child sexual abuse and child pornography online: not only regulatory frameworks have been provided in a timely manner, but the guidance provided has been consistent and constant, and has evolved considering the specific developments of these issues and attempting to readapt regulatory frameworks to address newly emerging concerns that were specifically linked to the online environment.

On the other hand, Facebook's approach to protection of children online has been insufficient, and its incorporation of international standards into their content policies has been slow, especially in light of the availability of guidance in this area. Facebook's Terms of Service were silent on measures aimed at the protection of children on the platform, with the sole exception of the implementation of age restrictions: Facebook policymakers may have the found it easiest to ban use of the platform to those under 13 with these instruments in mind. Taking into account the origins of the platforms, its impact at the time in terms of public discourse and the relatively limited user base, this might have been perceived as a reasonable approach. Nonetheless, as already mentioned above, already in 2007 the company faced backlash when the New York Times reported that a "concerned parent" had created a fake Facebook profile of a fifteen-year-old girl

[336] Maria Grazia Giammarinaro, 'Report of the Special Rapporteur on Trafficking in Persons, Especially Women and Children' (2014) A/69/269.

[337] Ibid.

who was "looking for trouble" to see how dangerous Facebook was.[338] Then New York Attorney General Cuomo put pressure on Facebook to increase its scrutiny of child grooming following the sting operation and, after a three-week negotiation, Facebook settled with New York: all reports of unwanted harassment or pornography had to be handled within 24 hours[339].

It is worth recalling that in this same year the Lanzarote Convention was adopted by the Council of Europe, which includes ICT companies in the category of relevant stakeholders that should participate in the development and implementation of policies to prevent sexual abuse and exploitation of children.

But it is only in 2013 that Facebook's Community Standards make a first reference to protection of minors, stating that Facebook has a strict policy against the sharing of pornographic content and "any explicitly sexual content where a minor is involved." At this point, Facebook could have already drawn from several instruments to draft a more detailed and specific standard. For example, Facebook could have borrowed the definitions that the Lanzarote Convention offered for child sexual abuse, child pornography, and child prostitution. Instead, Facebook uses the term "sexual content" as encompassing all these definitions. Similarly, Principle 7 of the Safer Social Networking Principles for the EU (2008) already required social networks to moderate content and identify risks to children and minors.

Subsequent updates to Facebook's policies on the protection of children appear to be slightly more aligned with international standards, with the 2017 version of the Community Standards offering a more detailed definition than the 2013 version. For example, the 2014 UN Report of the Special Rapporteur on the sale of children, child prostitution and child pornography, Maud de Boer-Buquicchio adds "grooming" to the list of forms of exploitation and abuse in the Optional Protocol to the Convention on the Rights of the Child on the sale of children, child prostitution and child pornography. This offence was already introduced at the EU regional level by the Lanzarote Convention, and it seemed to play into Facebook's definition of sexual exploitation by including

[338] Stone (n 335).

[339] 'Facebook Settles New York Child Safety Probe' (*Reuters*, 15 October 2007) https://www.reuters.com/article/us-facebook-settlement-idUSN1539478220071016.

the "solicitation of sexual material." Facebook also introduced the referral of content to law enforcement.

The 2020 Facebook's Community Standards on "Child Sexual Exploitation, Abuse and Nudity" represents a major improvement and largely reflects the content of international standards, indicating a strong level of compliance. Notably, the 2020 version the Community Standards bans "content (including photos, videos, real-world art, digital content and verbal depictions) that shows minors in a sexualized context," to which the UN Special Rapporteur on the sale and sexual exploitation of children, including child prostitution, child pornography and other child sexual abuse material had drawn attention to in her 2020 Report, identifying "drawings and virtual representations of non-existing children in a sexualized manner, widely available on the Internet" as a new trend that "appears to normalize [child sexual abuse] and may encourage potential offenders and increase the severity of abuse."[340] In the same Report, the Special Rapporteur articulated the importance of the ICT sector for prevention and response strategies in the context of child exploitation, underlining the importance of cooperating with law enforcement agencies. In this 2020 version of the Community Standards, Facebook also offers a more detailed description of how the company cooperates with law enforcement by stating that they report apparent child exploitation to the National Center for Missing and Exploited Children (NCMEC), in compliance with applicable law. In 2020 Facebook has also announced it has joined an industry initiative, Project Protect: A plan to combat online child sexual abuse, to fight child exploitation online.[341]

k) The tension between intellectual property, access to knowledge, and the exercise of freedom of speech on Facebook

[340] Boer-Buquicchio (n 188) para 63.

[341] Antigone Davis, 'Facebook Joins Industry Effort to Fight Child Exploitation Online' (*Facebook*, 11 June 2020) https://about.fb.com/news/2020/06/fighting-child-exploitation-online/.

Intellectual property is a domain with a documented history of tension with freedom of expression, particularly in the realm of copyright law.[342] Intellectual property grants exclusivity over the use of intellectual creations, and this can be used in ways that conflict with freedom of expression and other human rights that are inextricably linked to it under the prism of freedom of access to information.[343] For its part, Article 19 ICCPR ensures the right to freedom of expression, including the freedom to "…seek, receive and impart information…" Furthermore, Article 27 of the Universal Declaration of Human Rights provides for everyone's right to "freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits," which is also replicated in Article 15 of the International Covenant on Economic, Social, and Cultural Rights, recognizing the right of everyone to enjoy the benefits of scientific progress and its applications, while also protecting "the freedom indispensable for scientific research and creative activity."

Building upon these foundational pillars, UNESCO convened a meeting of experts in 2009 which resulted in the so-called "Venice Statement on the Right to Enjoy the Benefits of Scientific Progress and its Applications", recognizing *inter alia* that a human rights-based approach requires that science and its applications are consistent with fundamental human rights principles such as non-discrimination, gender equality, accountability and participation, and that particular attention should be paid to the needs of disadvantaged and marginalized groups. It also explicitly called the private sector to examine ways of contributing to this right by giving greater attention to the basic needs of disadvantaged and marginalized groups, and in particular the right of all to enjoy the benefits of scientific progress.

In 2012, the Annual Report of the UN Special Rapporteur in the field of cultural rights was devoted entirely to "The right to enjoy the benefits of scientific progress and its applications", recognizing among other things that "the rights to science and to culture should both be

---

[342] Alexandra Couto, 'Copyright and Freedom of Expression: A Philosophical Map' in Axel Gosseries, Alain Marciano, and Alain Strowel (eds), *Intellectual Property and Theories of Justice* (Palgrave Macmillan, 2008); Kembrew McLeod, *Freedom of Expression- Resistance and Repression in the Age of Intellectual Property* (University of Minnesota Press 2007).

[343] Frank La Rue, 'Report of the Special Rapporteur on the Promotion of Freedom of Opinion and Expression' (2011) A/HRC/17/27, paras 49-50.

understood as including a right to have access to and use information and communication and other technologies in self-determined and empowering ways."[344] In 2020, General Comment 25 on article 15 of the UN Committee on Economic, Social and Cultural Rights[345] clarified that the concept of "culture" is broader than science, including other aspects of human existence; and called States to make every effort, in their national regulations and in international agreements on intellectual property, to guarantee the social dimensions of intellectual property, in accordance with the international human rights obligations they have undertaken. In particular, it referred to the need to strike a balance between intellectual property and the open access and sharing of scientific knowledge and its applications.

In the same year, the UN Special Rapporteur on Freedom of Opinion and Expression published a report on Artistic Freedom of Expression,[346] where it recognized that despite the significant research analyzing private content moderation systems, artists are often wrongfully censored for posting material that is controversial or in any way subjectively offensive to any user. This is due primarily to problems of artificial intelligence: besides the race and sex biases that are incorporated in the design of the artificial intelligence, it has trouble with the intricacies of language, being unable to grasp the complexities of colloquial speech and humour.[347] The natural implication of that is the importance of allowing independent testing of existing mechanisms and technologies used for content moderation, which has often been invoked by researchers to conduct experiments on Facebook's properties.

Facebook's receptiveness to this kind of calls has been criticized, due to the prohibitions set in their terms to engage in activities that may be seen as constitutive elements of this kind of research. For instance, the Terms of Service in 2005 prohibited printing and reproduction of the material accessed on Facebook other than for personal, non-commercial use, the republication

---

[344] Farida Shaheed, 'Report of the Special Rapporteur in the Field of Cultural Rights' (2012) A/HRC/20/26.

[345] General comment No. 25 (2020) on science and economic, social and cultural rights (article 15 (1) (b), (2), (3) and (4) of the International Covenant on Economic, Social and Cultural Rights (2020) E/C.12/GC/25 para 10.

[346] David Kaye, 'Research Report on Artistic Freedom of Expression' (2020) A/HRC/44/49/Add.2.

[347] Ibid para 44.

of the website's content on any Internet, Intranet or Extranet site, the incorporation of the information in any other database or compilation and the modification, copying, distribution, framing, reproduction, republication, download, display, posted, transmission, or sale of the website's content, in whole or in part, without prior written permission. They also allowed Facebook to remove or restrict access to content, services or information if they determined that doing so is reasonably necessary to avoid or mitigate adverse legal or regulatory impacts to Facebook. Since June 2018, the Terms of Service also prohibited access to or collection of data from its own Products using automated means (without prior permission).

Facebook has occasionally invoked these terms as the legal basis to stop attempts to conduct research on important societal issues on the platform: this includes a high-profile case against Ad Observer, a browser add-on that aims to collect advertising data to improve the accountability of political ad campaigns.[348] After sending a cease-and-desist letter to the project administrators and an ensuing negotiation, in 2021 Facebook decided to shut down the account of the researchers who created the Ad Observer tool, citing privacy concerns and the need to comply with the consent decree between Facebook and the FTC since 2020.[349] This attracted significant public attention, in particular due to the fact that the consent decree does not *require* Facebook to shut down privacy-respecting research like the one of the Ad Observer (as pointed out in a letter to

[348] Laura Edelson and Damon McCoy, 'We Research Misinformation on Facebook. It Just Disabled Our Accounts' *New York Times* (10 August 2021) https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html.

[349] Mike Clark, 'Research Cannot Be the Justification for Compromising People's Privacy' (*Facebook*, 3 August 2021) https://about.fb.com/news/2021/08/research-cannot-be-the-justification-for-compromising-peoples-privacy/.

Mark Zuckerberg published by the FTC the next day),[350] even if one could argue that the decree *justifies* taking action to mitigate the risk of adverse legal or regulatory impact[351].

Regardless of the merit of this claim, the incident revitalized the discussion initiated by the Knight First Amendment Institute in 2018 with a letter to Facebook calling for the introduction of a safe harbour in Facebook's terms to for certain news-gathering and research projects.[352] The proposal, to which Facebook has not yet publicly responded, is that nobody would violate the Terms by collecting public information through automated means, or by creating or using temporary research accounts, as part of a news-gathering or research project, so long as a number of conditions are met.

While it must be conceded that Facebook remains a private enterprise and thus is in principle not required to give access to its internal data, the proposal would not specifically require that, nor demand substantive changes to the existing terms: it would simply involve a concession by Facebook not to enforce those terms in a way that outlaws automated data collection of public information and the creation and use of temporary research accounts, as long as they meet the specified conditions that ensure that the research is aimed to inform the general public about matters of public concern, follows adequate privacy and security standards, and does not mislead third parties.

Even leaving aside the context of permitting research access, it is clear that intellectual property has been a prominent concern from the very beginning of Facebook's history, and one that has clear implications for freedom of expression: enforcement of intellectual property, and

---

[350] Samuel Levine, 'Letter from Acting Director of the Bureau of Consumer Protection to Facebook' (*FTC*, 5 August 2021) https://www.ftc.gov/news-events/blogs/consumer-blog/2021/08/letter-acting-director-bureau-consumer-protection-samuel.

[351] James Vincent, 'Facebook's justification for banning third-party researchers 'inaccurate,' says FTC' (The Verge, 6 August 2021) https://www.theverge.com/2021/8/6/22612545/facebook-banned-third-party-researchers-inaccurate-says-ftc

[352] Knight First Amendment Institute at Columbia University, 'Knight Institute Calls on Facebook to Lift Restrictions on Digital Journalism and Research' (Columbia University, 7 August 2018) https://knightcolumbia.org/content/knight-institute-calls-facebook-lift-restrictions-digital-journalism-and-research.

particularly copyrights and trademark, may give rise to significant restrictions of the free flow of information if defenses and exceptions are not given an appropriate role at the enforcement stage. In this regard, it is important to note that, throughout the years, Facebook's policies shifted towards a more permissive approach to the sharing and reproduction of content that is posted on the site, conceivably as a reflection of the maturation of the understanding of the strategic role of cross-posting to the company's business model, and the evolution to a more user-centric approach to enforcement.

To give a vivid illustration of that, one needs to look no further than the first Terms of Service (version of November 2005), which included restrictions on users' ability to reproduce the content of the site without Facebook's prior written permission, to download and print a copy of that content for non-personal and commercial use, and to re-publish content "on any Internet, Intranet or Extranet site or incorporate the information in any other database or compilation." That initial policy changed in 2007 (Terms of Service version of June 2007), when the attention begun to shift toward the possible copyright violations with regard to the content posted by users on the site, requiring them to refrain from making available videos other than those of personal nature either depicting the user or her friends, or taken by that user or her friends, or otherwise that constitute original art or animation created by the user or her friends. In 2009 (Terms of Service version of June 2009), Facebook introduced into its terms the granting of a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that users post on or in connection with Facebook, causing significant backlash due to the possibility it creates for the re-use of user content for unlimited purposes by Facebook and its commercial partners.[353]

Facebook's approach towards IP shifted from a focus on protecting Facebook's intellectual assets to one of support for the intellectual property of third parties, while simultaneously requiring users to take responsibility for the content they make available. However, at the same time, it can be criticized that

[353] Catherine Lyons, 'Facebook Can Use Your pictures for Ads, No Permission Required' *LA Times* (24 July 2009).

Facebook's terms remain silent on the use of automated technologies for the detection and removal of potentially infringing content.

Interestingly, this clause was removed with the revision of the terms in 2011 (Terms of Service version of April 2011), replaced since 2017 in the Community Standards (version of January 2017) by a recognition that users own all the content posted on Facebook, but combined since the 2018 terms (Terms of Service version of Jan 2018) with a compulsory license to grant Facebook permission to store, copy, and share it with others (consistent with user settings) such as service providers that support Facebook service or other Facebook Products used by that individual. At the same time, since 2017 Facebook begun requiring users who intend to post content to first verify that they have the right to do so, referring to copyright, trademarks and other legal rights, and declared its commitment to helping people and organizations promote and protect their intellectual property rights, violations of which are not allowed on the site.

On the basis of this, we can observe that Facebook's approach towards IP shifted from a focus on protecting Facebook's intellectual assets to one of support for the intellectual property of third parties, while simultaneously requiring users to take responsibility for the content they make available. From a freedom of expression standpoint, this denotes more openness in the re-use of Facebook's assets, by eliminating prior restraints on speech imposed by Facebook, such as the need to obtain a permission.

However, at the same time, it can be criticized that Facebook's terms remain silent on the use of automated technologies for the detection and removal of potentially infringing content, which practically speaking appears to be an inevitable measure in response to the recent legislation introduced in the European Union to deal with copyright in the digital single market: as discussed in Part 1, the Directive requires online content sharing providers who wish to escape liability for third-party content to make best efforts to obtain an authorization from the rightsholders, to make "best efforts to ensure the unavailability of specific works and other subject matter for which the rightsholders have provided the necessary information," and in any event "act expeditiously, upon receiving sufficiently substantiated notice from the rightsholders, to disable access or remove from their website the notified work or other subject matter, and [make] best efforts to prevent

their future uploads […]".[354] In this regard, it bears noting that in 2019 the UN Special Rapporteur on Freedom of Opinion and Expression, David Kaye, publicly criticized this article of the draft Directive as it appeared "destined to drive internet platforms toward monitoring and restriction of user-generated content even at the point of upload", which he saw as "sweeping pressure for pre-publication filtering" that is "neither a necessary nor proportionate response to copyright infringement online."[355] He specifically criticized the absence of specific requirements on platforms and member states to defend freedom of expression, which makes it unclear how either will comply with the Directive's proposed safeguards, such as the requirement that "quotation, criticism, review" and the "use [of copyrighted works] for the purpose of caricature, parody or pastiche" be protected. He also added the preoccupation that misplaced confidence in filtering technologies to make nuanced distinctions between copyright violations and legitimate uses of protected material would escalate the risk of error and censorship. Given these concerns, it would have been desirable for Facebook to specifically incorporate safeguards for freedom of expression into its terms relating to the implementation of this provision.

l) A venerable yet failed experiment in digital democracy: Governance & stakeholder involvement in shaping free speech on Facebook

The way in which Facebook opens itself to feedback by its community of users in the process of introducing potentially consequential new terms forms part of the framework that governs freedom of expression on its platform, as users exercise their rights within the confines of Facebook's norms. In this area, Facebook understood from its first few years of operation that it was important to empower its community, recognizing at least to some extent their right to self-determination, and embarked in two efforts to gather feedback in potentially complementary ways.

---

[354] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, OJ L 130, art 17(4).

[355] David Kaye, 'EU Must Align Copyright Reform with International Human Rights Standards' (*UN*, 11 March 2019) https://www.ohchr.org/en/NewsEvents/Pages/DisplayNews.aspx?NewsID=24298.

In 2009, following backlash regarding the introduction of new privacy rules on the site,[356] Facebook introduced a new rule in its terms that was specifically designed not only to allow such feedback, but also to make it binding on Facebook. The rule created a notice and comment process by which Facebook would give users the opportunity to participate in a vote regarding a proposed change of terms of service if more than 7,000 users commented on it. The rule also established that the vote, which involved a choice between alternative versions of the new terms, would be binding on Facebook if more than 30% of all active registered users as of the date of the notice voted. In announcing the rule, Mark Zuckerberg placed great emphasis on the role of public input and engagement for developing content moderation policies for enabling companies "to consider the human rights impact of their activities from diverse perspectives,"[357] explaining that companies like Facebook need to develop new models of governance.

Indeed, the introduction of such mechanism was unprecedented, even though the right to self-determination of online communities had been proclaimed and proposed as a legal concept in the early cyberlaw literature,[358] and even if the involvement of stakeholders in drafting terms had been suggested by the 2018 CoE Recommendation on the Role and Responsibilities of Internet Intermediaries, according to which "the process of drafting and applying terms of service agreements, community standards and content-restriction policies should be transparent, accountable and inclusive. Intermediaries should seek to collaborate and negotiate with consumer associations, human rights advocates and other organisations representing the interests of users and affected parties, as well as with data protection authorities before adopting and modifying

[356] Adi Robertson, 'Mark Zukerberg Wants to Democratize Facebook-here's What Happened When He Tried' (*The Verge*, 05 April 2018) https://www.theverge.com/2018/4/5/17176834/mark-zuckerberg-facebook-democracy-governance-vote-failure.

[357] Facebook, 'Facebook Opens Governance of Service and Policy Process to Users' (*Facebook*, 26 February 2009) https://about.fb.com/news/2009/02/facebook-opens-governance-of-service-and-policy-process-to-users/.

[358] John Perry Barlow, 'A Declaration of the Independence of Cyberspace' (*Electronic Frontier Foundation*, 8 February 1996) https://www.eff.org/pt-br/cyberspace-independence; David R. Johnson and David Post, 'Law and Borders: The Rise of Law in Cyberspace' (1996) 48 Stanford Law Review 1367.

their policies. Intermediaries should seek to empower their users to engage in processes of evaluating, reviewing and revising, where appropriate, intermediaries' policies and practices".[359]

In conjunction with this new policy, Facebook rolled out two documents that could be seen as foundational for a process akin to a privately-led "constitutionalization" of the rules of the site, specifically regarding the values that ought to inform the development of the service and on Facebook's commitments related to it: the Principles and the Statement of Rights and Responsibilities. People were invited to comment on these documents by joining dedicated groups (today, these groups are closed and the pages are no longer available),[360] and subsequently vote on alternative versions that Facebook made available on the site. The voting on these two initial documents was automatically opened, differently from future changes that would require more than 7000 comments on Facebook´s governance page.[361]

While this innovative consultation process was welcomed by various stakeholders, its role in effectively putting a check against Facebook's adoption of controversial changes remained essentially null. First of all, a cursory look through the governance page reveals that Facebook never reached 7000 comments on a post, which suggests that the company was either too optimistic regarding user participation in its governance process, or strategically chose this number to make this a mechanism that would rarely bite. Secondly, the requirement of 30% turnout over Facebook´s active user base was equally, or perhaps even more, unrealistic, particularly in the absence of a rigorous process of notification for all users in ways that are effective considering the different linguistic and educational backgrounds.[362]

---

[359] See Appendix to Recommendation CM/Rec(2018)2, Guidelines for States on Actions to Be Taken Vis-À-Vis Internet Intermediaries With Due Regard to Their Roles And Responsibilities, para 2.2.2.

[360] Respectively at: http://www.facebook.com/group.php?gid=54964476066 and http://www.facebook.com/group.php?gid=67758697570.

[361] Facebook Site Governance https://www.facebook.com/fbsitegovernance.

[362] Anita Ramasastry, 'The Failure of Facebook's Voter Experiment: What It May Mean' (*FindLaw*, 7 May 2009) https://supreme.findlaw.com/legal-commentary/the-failure-of-facebooks-voter-experiment-what-it-may-mean.html.

In announcing the results of the voting of the two founding documents, Facebook stated that it made significant efforts to make voting easy and to give everyone the opportunity to vote — including by translating the documents and voting application into several of the most popular languages (but not all of them) on the site, showing a message about the vote on users' home pages, and running advertisements and videos across Facebook promoting the vote,[363] but provided no further details to describe each of those steps. This resulted in a very low participation threshold (665,654 votes, equivalent to 0.3% of active users) and a 74.37% approval rate of those documents.

Facebook's first reaction was that they would consider lowering the 30% threshold for future votes,[364] but ultimately it maintained that threshold until it published a proposed change that would get rid of voting in 2012,[365] explaining that the goal was "to make sure that we receive feedback from you in the best, most productive way possible so that we can be responsive to your input."[366] This proposed change was overwhelmingly rejected, but once again, the vote could not be made binding due to the low level of participation.[367] In other words, the direct democracy experiment failed, and at least in part, this could be attributed to the lack of an effective way to involve users in the process: it is arguable that the result would have been different if Facebook, for instance, made use of the site conditional on scrolling through the proposed changes explained in simple terms, and a final option to vote with a click.

It is also worth noting that Facebook did not follow up on the announcement made when it rolled out this new governance rules, when it said it had the intention of establishing a user council

---

[363] Ted Ullyot, 'Results of the Inaugural Facebook Site Governance Vote' (*Facebook*, 23 April 2009) http://web.archive.org/web/20090430215524/http://blog.facebook.com/blog.php?post=79146552130.

[364] Ibid.

[365] Kimber Streams, 'Facebook Proposes Policy Changes, Will Share User Data with Instagram and Kill User Veto' (*The Verge*, 21 November 2012) https://www.theverge.com/2012/11/21/3676518/facebook-data-use-instagram-filters-vote.

[366] Adi Robertson, 'Mark Zukerberg Wants to Democratize Facebook-Here's What Happened When He Tried' *(The Verge*, 05 April 2018) https://www.theverge.com/2018/4/5/17176834/mark-zuckerberg-facebook-democracy-governance-vote-failure.

[367] Ibid.

to participate more closely in the development and discussion of policies and practices. While this process began by inviting the authors of the most insightful and constructive comments on the draft documents to serve as founding members of the groups where the changes were voted,[368] no further steps were publicly announced regarding the creation of the council.

Instead, Facebook veered towards the use of a different mechanism to gather outside views on the policies, by making explicit that the company seeks to engage experts, in addition to Facebook users. Updates of its terms in November 2020 included a clause that announced the creation of the Stakeholder Engagement team, a sub-team of Product Policy, "whose main goal is to ensure that Facebook's policy development process is informed by the views of outside experts and the people who use Facebook." It also made explicit that specific practices and a structure for engagement were developed in the context of the Community Standards, and that this would be expanded to cover additional policies, particularly ads policies and major News Feed ranking changes. As a general principle, the work of the Stakeholder Engagement team frames up policy questions requiring feedback and determines what types of stakeholders to prioritize for engagement. Facebook then reaches out to external stakeholders, gathering feedback that they document and synthesize internally for the company. No specific form of consultation is prescribed, however, and Facebook recognizes that "the heart of [their] approach to engagement" is private conversations, typically without releasing the names of people involved in order to stimulate open engagement and protect confidentiality[369]. The update also mentions that sometimes group discussions are convened around particular regions or policy areas, and occasionally Facebook reaches out to relevant Facebook users to get their view (without, however, explaining how this concretely works).

One could raise criticism about the transparency and accountability of this process, as it does not allow to identify the sources of the input received. Nevertheless, it is remarkable that Facebook commits to consulting about the pros and cons of proposed changes and to present at the Product Policy Forum a detailed summary of the feedback received, together with a spectrum of policy options, both of which are made public and followed by an announcement of the decision taken

---

[368] Ullyot (n 363).

[369] Facebook, 'Stakeholder Engagement' https://www.facebook.com/communitystandards/stakeholder_engagement

and the rationale for it. Furthermore, Facebook publicly commits in its terms on working together with BSR on ways to improve the involvement of stakeholders (not just users) in the design of products and services, including through the use of questionnaires and free-form questions. This aligns with one part of the specific CoE Recommendation on the Role and Responsibilities of Internet Intermediaries of 2018, which emphasises the importance of involving experts and stakeholders, but the repeal of the user voting provision may be seen as a relinquishment of an important effort to involve users in the drafting of new terms. Although it must be conceded that there may be overlap between the groups of stakeholders and users, the learning from the direct democracy attempt suggests that it will be crucial to think about ways in which the communication and use of these opportunities for participation can be made effective.

m) The danger of over-reliance on automatic content moderation: automated technologies and their impact on regulating freedom of expression on Facebook

While the legislative requirements introduced by EU copyright legislation offer a clear example of the battlefield regarding automated content moderation and freedom of expression, the criticism around the accuracy of algorithmic content moderation runs wider than that: the use of automated technologies can raise obstacles to freedom of expression and access to knowledge across the board, including literally all the areas of content moderation mentioned in this report. To address these concerns, calls have been made for platforms to disclose the rate of false positives and false negatives, and to undertake a thorough, transparent and independent review of the implications of automation.[370] In particular, the Report of the UN Special Rapporteur on Freedom of Expression in 2018 recommended the publication of granular data on content removals and contestation thereof, the engagement in human rights impact assessments, as well as the

[370] See for instance Richard Ashby Wilson and Molly K. Land, 'Hate Speech on Social Media: Content Moderation in Context' (2021) 52 Connecticut Law Review 1029.

ongoing monitoring and auditing over the use of AI.[371] Additionally, the Report recommended that companies make explicit where and how AI technologies are used in their platforms, services and applications, and use innovative ways to signal to individuals when they are subject to an AI-driven decision-making process, when AI plays a role in displaying or moderating content and when their personal data may be integrated into a dataset that will be used to inform AI systems.[372]

A related concern has to do with the opacity of content moderation as applied in individual cases. For instance, Suzor et al. note widespread confusion among platform users on the exact content or behaviour that triggered a sanction and how it came under review, as well as a systemic failure by platforms to provide good reasons to explain their content moderation decisions.[373] This is in line with the results of a small-scale empirical study,[374] which also documents a varying degree of rigidity in the application of content moderation rules across different categories of content, in some ways inconsistent with the findings of Facebook's Community Standards Enforcement Report.[375] To improve transparency and accountability of content moderation,

---

[371] David Kaye, 'Report on Artificial Intelligence Technologies and Implications for Freedom of Expression and the Information Environment' (2018) A/73/348, paras 67-68.

[372] Ibid para 66.

[373] Nicolas P. Suzor, Sarah Myers West, Andrew Quodling, and Jillian York, 'What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation' (2019) 13 International Journal of Communications (online).

[374] Nicolo Zingales et al., 'Report of Field Project: Content Moderation & Freedom of Expression @ Facebook' (2020) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3897035. The study presented the following figures in relation to content moderation occurred between February and May 2020: figures in relation to content moderation occurred between February and May 2020:

- Nudity and sexual activity: 2/22 instances of content moderated proactively, none reactively.
- Hate speech: 0/3 instances of content moderated proactively, 1 reactively;
- Bullying and harassment: 0/14 instances of content moderated proactively, none reactively.
- Regulated goods and firearms: 2/2 instances of content moderated proactively, none reactively.
- Violent and graphic content: 6/43 instances of content moderated proactively, 5 reactively.
- Dangerous organizations: 0/17 instances of content moderated proactively, none reactively.
- Spam: 2/3 instances of content moderated proactively, none reactively.
- Suicide & self-injury: 2/21 instances of content moderated proactively, none reactively.

[375] See 'Facebook Community Standards Enforcement Report' (May 2020) https://about.fb.com/news/2020/05/community-standards-enforcement-report-may-2020/>, which presented the following statistics in relation to January-March 2020:

- Nudity and sexual activity: 39.5 million instances of content acted upon, 99.2 % proactive;
- Hate speech: 9.6 M acted upon, 88, 8% proactive;
- Bullying and harassment: 2.3 million acted upon, 15.6% proactive;

therefore, the study highlighted the need to combine disclosure of aggregate numbers with an *independent assessment* over the existence of prohibited content, as the currently used measure of prevalence of prohibited content available on the site may otherwise produce partial and oversized figures.[376]

The importance of information about content moderation as applied in individual cases plays a central role also within the broader set of principles on transparency and accountability that have been offered by a group of private sector organizations, academic experts and civil society advocates on the occasion of the first Content Moderation at Scale conference in Santa Clara, CA on February 2nd, 2018.[377] These principles, which have become known as the Santa Clara Principles on Transparency and Accountability on Content Moderation (or simply, the Santa Clara Principles) focused on three cornerstones:

- Transparency: companies should publish the numbers of posts removed and accounts (temporarily or permanently) suspended due to violation of content guidelines.
- Notice: companies should provide notice to any users whose content is taken down or account is suspended about the reasons for doing so.
- Appeal: companies should provide a meaningful opportunity for a timely appeal of any content removal or account suspension.

Recently, a call for submissions was published for an update of the principles to include specific recommendations for transparency around the use of automated tools and decision-making,

---

- Regulated goods and firearms: regulated goods 7.9 million acted upon, 99.1 % proactive; firearms 1.4 million acted upon, 71% proactive;
- Violent and graphic content: 2.3 million acted upon, 94.7% proactive;
- Dangerous organizations: organized hate 4.7 million, acted upon, 96.7% proactive; terrorism 6.3 million acted upon, 99.3% proactive;
- Spam: 1.98 million acted upon, 99.9% proactive;
- Suicide & self-injury: 1.7 million acted upon, 97.7 % proactive;
- Fake accounts: 17 million acted upon, 99.7% proactive.

[376] Guy Rosen, 'Community Standards Enforcement Report' (*Facebook*, November 2020) https://about.fb.com/news/2020/11/community-standards-enforcement-report-nov-2020/.

[377] Ibid.

among other issues.[378] This has led to the formulation of new demands for transparency in the Santa Clara Principles 2.0[379], including: when and how automated processes are used (whether alone or with human oversight) when actioning content; the categories and types of content where automated processes are used; the key criteria used by automated processes for making decisions; the confidence/accuracy/success rates of automated processes, including changes over time and differences between languages and content categories; the extent to which there is human oversight over any automated processes, including the ability of users to seek human review of any automated content moderation decisions; the number (or percentage) of successful and unsuccessful appeals when the content or account was first flagged by automated detection, broken down by content format and category of violation; and participation in cross-industry hash-sharing databases or other initiatives and how the company responds to content flagged through such initiatives. In light of this, a commitment by Facebook to adopt suitable measures addressing the aforementioned concerns could well be expected, particularly in light of the fact that Facebook itself recognizes that artificial intelligence plays an increasingly larger role in content review.

7. Recommendations

- **A content moderation acquis and a board to enforce it.** Facebook's content moderation policies have improved vastly since its creation in 2004. The significantly more detailed Community Standards today delimit Facebook's control over users' right to freedom of expression much more robustly than the laconic Terms of Service when Facebook was first launched. The progress that has been achieved is to Facebook's credit, but it also represents a series of victories for users. We recommend that this *acquis* of freedom of expression rights be recognized and maintained. The recognition of a freedom of expression acquis on Facebook would suggest that, going forward, Facebook cannot amend its Terms of Service and Community Guidelines in ways that result in the

[378] The Santa Clara Principles, 'Call for Submissions' (March 2020) https://santaclaraprinciples.org/cfp/.

[379]

undue curtailment of users' freedom of expression rights beyond current limitations. The aim is to ensure that the protection of freedom of expression on Facebook can only get better from here. We envisage a board similar to the Oversight Board that will review proposed changes to Terms of Service and Community Standards and reject those that it deems infringe on the acquis. We do not purport this to be an easy task; our report documents numerous instances where freedom of expression restrictions were necessary to protect conflicting rights within the limitations prescribed by freedom of expression standards, and the proposed board will have to engage in such balancing. The proposed acquis does not imply that no further clarifications or limitations can be prescribed in Facebook's content policies, but rather that they are reviewed and deemed necessary according to freedom of expression standards first, so that they are not unnecessary or disproportionate. The concept of acquis is well-recognized in various legal systems and contexts, including in human rights.[380] Most notably, it is associated with EU Law, where it incorporates "the fundamental principles concerning the structure of the legal order and the case law on the essential requirements of the Community, constitutes an untouchable hard core, that is, an absolute substantial restriction implicitly imposed on any revision."[381] Similarly, it is this "hard core" of freedom of expression protections that we see the acquis safeguarding.

- **A general application of the proportionality standard, particularly where detailed guidance is missing.** The prescriptive capacity of rules is non-exhaustive by nature. A look at the international instruments on freedom of expression and on Facebook's Terms of Service and Community Standards reveals that no matter how detailed they become, there will always be unregulated gaps. The 'legality, necessity, proportionality' standard tells us that, because the right to express oneself freely is the norm, any limitations should pursue a clearly stated legitimate aim (legal), they should be necessary to achieve that aim (necessary), and they should not go beyond what is required (proportionate). Where

[380] Sophie Robin-Olivier, 'Fundamental Rights as a New Frame: Displacing the Acquis' (2018) 14 European Constitutional Law Review 96.

[381] Carlo Curti Gialdino, 'Some Reflections on the Acquis Communautaire' (1995) 32 Common Market Law Review 1108.

Facebook provides detailed guidance in terms of the limitations that apply to different types of speech, there is an expectation that they reflect the legality, necessity, proportionality triptych and Facebook's policies include several instances of such balancing. Facebook's Community Standards, for instance, prohibit the celebration of committed crimes, but explicitly allow people to debate or advocate the legality of criminal activities, as well as address them in a humorous way. At the same time, the less detailed Facebook's restrictions are, the more imperative it becomes that they be applied within a proportionality framework, so as to limit arbitrariness, excessiveness, and unfairness. For instance, in the context of copyright, the fact that Facebook's terms do not mention the role of defenses like parody, satire, educational use and others may be perceived as an indication that those values are not adequately considered in enforcing Facebook's rules. The explicit acknowledgment of the preservation of those values would align with the specific recommendation made by the UN Special Rapporteur on Freedom of Expression in 2017 to reform and strengthen pre-existing policies and mechanisms to address violations of freedom of expression,[382] but in lack of such explicit mention, a proportionality standard could allow the scope for such exceptions even if they are not included in writing in Facebook's policies. The proportionality standard is universal in human rights law and asks that rules and sanctions be applied not in a black and white fashion but rather only to the extent that is necessary to achieve the stated goal. We recommend that Facebook adopt in its Terms of Service and Community Standards a general provision that it will apply the rules contained therein proportionately to the goal that each rule aims to achieve. Our proposal thus mirrors the recently introduced provision in the proposed Digital Services Act (Article 12(2)) which states that "providers of intermediary services shall act in a diligent, objective and proportionate manner in applying and enforcing the restrictions referred to in paragraph 1, with due regard to the rights and legitimate interests of all

[382] Although the recommendation is directed to Internet access providers and is presumably referring to changes addressing established violations of freedom of expression, there appears to be no reason why the gist of this advice cannot be exported to digital intermediaries more broadly as part of human rights due diligence: ToS and CS should not merely provide legal grounds for content moderation, but also an opportunity to commit to the respect the principles of freedom of expression that are recognized as part of international law.

parties involved, including the applicable fundamental rights of the recipients of the service as enshrined in the Charter." The inclusion of a proportionality standard in the Terms of Service will signify stronger commitment than simply complying with the DSA and make it more actionable on the side of Facebook's users. The proportionality principle should permeate the entirety of Facebook's Terms of Service and Community Standards as well as the implementation measures Facebook puts in place (e.g. the enforcement guidelines used by Facebook's content moderation staff).

- **An improvement on users' access to remedies.** Disputes between Facebook and users arise frequently and across all areas of expression. The current appeals process that Facebook offers its users, through which they can request the review of a decision related to their content, is not fully spelled out in Facebook's Terms of Service. The Terms of Service, in fact, currently make a mere reference to "let[ing users] know and explain any options that [users] have to request another review." While the procedure is explained in the Newsroom,[383] Facebook should include a description of this appeal procedure in the Terms of Service, which constitute a binding document between the company and the users, in order to improve the legitimacy, accessibility and predictability of the mechanism. Moreover, users should be offered the opportunity to explain the reasons why they are challenging Facebook's decision via a written statement. Such an improvement would not only bring this appeal process more in line with principles of procedural justice, but it would also widen users' access to remedies, especially in light of the limited number of cases that the Oversight Board can review. Additionally, allowing users to submit a written statement on the issue would not only make content reviewers better positioned to assess the context of the speech act, but, as also envisioned in the UNGPs, it would render the appeals mechanism "a source of continuous learning," since the feedback received by users would be beneficial for preventing future grievances and harm.

[383] Monika Bickert, 'Publishing our Internal Enforcement Guidelines and Expanding Our Appeals Process' (*Facebook*, 24 April 2018) https://about.fb.com/news/2018/04/comprehensive-community-standards/.

- **An explanation of whether and how artificial intelligence is used to enforce Terms of Service and Community Standards.** Automatic content moderation tools are widely deployed and are essential in maintaining social network quality. At the same time, they deny people the benefit of a human judgement. Facebook should commit in its Terms of Service/Community Standards to providing an illustration of the mechanisms applied to prevent overbroad and unfair application of these technologies, and to publishing granular data about their performance in content moderation. Users should also be explicitly conferred a right to be informed about AI-driven adverse decisions, a right to receive an explanation for such decisions, and the ability to contest them with the involvement of a human reviewer[384]. Once again, the proposed DSA goes in the right direction by asking hosting providers who remove or disable access to content to inform those who provided the content about that decision and its reasons, including information on the use made of automated means in taking the decision (art. 15 (2)(c)); as well as to refrain from taking solely automated decisions in their internal complaint handling mechanisms (art 17 (5)). However, it does not extend these duties to content moderation decisions other than removal or disabling of access, such as downgrading, disabling of comments or suspension of earnings, which can have a significant impact on the user's ability to reach audiences.
- **Scope for a bona-fide exception for the use of non-authentic names.** Facebook should change its real name policy to allow users to provide reasons as to why pseudonymity would be necessary for their activity on-site. The Terms of Service should also detail the criteria on the basis of which such requests are to be assessed, duly account for fundamental rights, and allow affected users to challenge adverse decisions before an independent third party.

[384] These requirements correspond to those imposed by the General Data Protection Regulation with regard to decisions based on automated processing of personal data that generate legal or similarly significant effects on data subjects (see article 22). However, our proposed rule would apply also when such decisions do not involve the processing of personal data, and having it enshrined in Facebook's terms would increase the user's awareness of these obligations.